LINKING RESEARCH, POLICY AND PRACTICE

WARSAW SCHOOL OF ECONOMICS

Easy access to large datasets with pivot tables in Excel & Tableau

Paper presented in track 9, "Institutional Research: working for the academic community" at the

EAIR 33rd Annual Forum in Warsaw, Poland

28-31 August 2011

Name of Author(s)

Keith Fortowsky
Kate McGovern

Contact Details

Keith Fortowsky, Director of Institutional Research
Office of Resource Planning, University of Regina
3737 Wascana Parkway
Regina, Saskatchewan, S4S 0A2
CANADA
E-mail: keith.fortowsky@uregina.ca

Key words

Institutional performance measures, Research design and methodologies, Information dissemination, Analysing data, Reporting data.

Abstract

Easy access to large datasets with pivot tables in Excel & Tableau

"Pivot Tables" are a powerful way for individual analysts to quickly create cross-tabulations and similar analysis of very large datasets. They are useful both for rapid ad hoc analysis and for building detailed reports that can be easily "filtered" by end-users to show only needed portions of a large dataset. Reports based upon pivot tables allow much analytic reporting to be moved from IT staff directly to analysts who know the actual business questions. We will demonstrate the use of pivot tables in both Microsoft Excel and a new product, Tableau. This presentation is intended primarily for those completely new to pivot tables, and secondarily for those new to Tableau. Examples focus on analysis of student diversity and retention.

Presentation

Easy access to large datasets with pivot tables in Excel & Tableau

Pivot tables are probably the biggest single productivity booster for individual analysts working with large data sets. We will demonstrate pivot tables using both Microsoft Excel and a relatively new product, Tableau. The presentation is intended primarily for those completely new to pivot tables, and secondarily for those new to Tableau.

Pivot tables are a highly 'tactile' tool in that data analyses can be created and rearranged almost instantly, through 'drag and drop', by an experienced user. It is much easier for a new user to grasp the fundamentals of 'pivoting' through a visual presentation than by reading about it. Tableau adds a whole new level of graphical functionality to pivot tables. Again, this is much easier to understand through visual demonstration.

This paper is intended to supplement, with background concepts and discussion, a visual demonstration of pivot tables. It provides an overview of the history of pivot tables. It highlights key concepts and terminology for use of pivot tables in MS Excel and Tableau. It also briefly examines the growing usage of Excel and Tableau pivot tables in the area of Institutional Research.

Since concrete examples are generally more accessible than abstract discussions, we will draw upon actual analysis work to illustrate some of the concepts discussed in this paper. The primary examples in the presentation will draw from student demographic and retention analysis as conducted at our medium-sized Canadian University. The results will ultimately be publicly available, and thus the data in this presentation is broadly realistic although not yet exact. There is, of course, no information that could be linked to individual students.

This paper will also discuss the limitations of Pivot Tables, which will be further illustrated in the presentation though examples from an analysis that utilizes both Excel and Tableau products.

History and functionality of PivotTables

Pivot table functionality has been included directly in Microsoft Excel (called "PivotTable", i.e. one word) since the mid 1990's and is thus easily available to any organization that uses Excel. Excel 2007 introduced a redesigned PivotTable interface that is significantly easier to use. If looking for books or other references, it is important to know if you are using a pre-2007 or post-2007 version of Excel.

Similar functionality is available in all other spreadsheet programs, including the open-source OpenOffice Calc (though with considerably fewer features). Many other programs also contain pivot table functionality, though often by another name (such as "cross-tab tables", "data pilot" in Calc; and the generic "pivot table" with a space). Pivot tables are implemented in most Report writers (ex. a "Cross-Tab Report" in Crystal Reports), statistical tools (SPSS, SAS) and hybrid tools such as dashboard toolsets.

A recent product, Tableau, makes the creation and distribution of highly visual pivot tables much easier than in just Excel. And the free "Tableau Reader" allows end-users to filter data and cut-and-paste results, creating a highly cost-effective reporting solution (www.tableausoftware.com).

MS Access and many SQL databases allow pivoting directly in the database, but the implementations are clumsy; exporting the data to another product is generally preferred, but the in-database functions can be useful for rapid initial checks and summaries of datasets. Pivoting is not a standardized SQL function and so the implementations have varying names, including "crosstab query", "pivot query" and "matrix query".

Most Business Intelligence platforms (ex. Cognos, SAS, Oracle BI) provide not only management of a large dataset (typically in a “data warehouse”) but also provide versions of the tools described above, intended for end users. However these tools are generally both expensive on a per user basis and, despite claims of ease-of-use, difficult to use. For both these reasons, the BI tools are often confined to a small group of “super users."

Posting on a website greatly enhances an institution’s ability to not only disseminate valuable information, but also allows the more informed user to further explore datasets to meet specific needs. Pivot tables offer a variety of options for posting to a website; we will briefly discuss some of these options in our presentation.

Use of PivotTables in Institutional Research

Advances in both data storage capabilities and analytical tools have enhanced the ability of Institutional Research (IR) Offices to better meet internal and external demands for information about the university, as well as broader post-secondary trends and issues. In light of the sheer volume of data that institutions are now capable of producing, combined with ever increasing demands for quickly providing current and reliable data, IR analysts are continually searching for better methods and tools.

IR analysts rely on a multitude of tools to fulfil analytical and reporting requirements. Pivot table capabilities, in particular, have gained growing interest by the Institutional Research and Analysis Community in North America since 2000. It has become a common feature in workshops and paper presentations at National and Regional IR forums in Canada and the United States. Prior to 2008, most workshops offered introductory sessions geared towards users who are familiar with Excel, but unfamiliar with pivot tables. These workshops promised to teach users how to generate a wide range of reports in far less time than if traditional methods were used. In the past four years forums have started to offer intermediate and advanced Excel PivotTable training sessions, as well as an increasing number of paper sessions describing new and novel features and uses. Since 2009, Tableau workshops and paper session have began to appear in the program, with presenters promoting this new product as an alternative or a complement to Excel on the basis of its ease of use, and exceptional visual analytical capabilities.

A core attraction of these products to IR Offices is the ease and speed in which a user can perform rapid ad hoc analysis of very large complex datasets. The key difference between pivot tables and conventional "static" reports is that a user can change the pivot table’s structure in seconds, by dragging and dropping fields to create or modify rows, columns, and the data to be summarized by these rows and columns. Further changes can be made by ‘filtering’ to leave only specific subsets of data. Pivot Tables that are already set up can be intuitive and easy to use, so with a small bit of training, “power users” can quickly become highly proficient and greatly lessen requests for custom data extractions and reports.

At the University of Regina (in Saskatchewan, Canada), the Office of Resource Planning (ORP) is the primary unit responsible for providing the executive officers and academic administrative units throughout the university with management information and professional services to support data access, planning and effective decision-making. ORP also provides information about the university to government and other external organizations.

Like many other institutions in North America, the ORP at the University of Regina meets its key reporting and data analysis functions using multiple software products. Many of the institution's standard reports used annually for planning purposes (e.g., Factbook reports), or

those used more frequently (e.g., daily or weekly) for management purposes are derived from pre-built Access database report queries. Over the past few years, ORP has started to migrate many of these reports into interactive Excel PivotTable reports. Most of the new reports developed by ORP and virtually all ad hoc reports are either developed from Excel PivotTables or using Tableau software, and increasingly these are distributed as interactive pivot tables or charts. At present ORP uses Tableau Desktop primarily as an internal analytical and data quality assessment tool, and to a limited extent for preparing static reports for end-users. The Office's long-term plan is to make institutional reports available to users via Tableau.

Retention Analysis at the University of Regina: Using Tableau and Excel to generate knowledge

At the University of Regina monitoring of student retention and completion is considered an integral part a sound strategic enrolment management strategy. Research and analysis in this area supports academic program reviews as well as campus-wide retention initiatives. Further retention work undertaken by ORP is contributing to a new initiative, the Western Canadian Universities Dataset (WCUD). This initiative was advanced in 2010 by senior administrators composed primarily of Vice Presidents Academic (VPAs) from Western Canadian Universities. Its purpose is to provide a rich regional database of comparative student information to help inform and foster discussion on issues regarding post-secondary education and trends.

WCUD bears some resemblance to another more comprehensive regional university database program that was implemented in Ontario in about 2006, Common University Data Ontario (CUDO). Both offer key institutional data at an "academic program" level of detail, organized according to common definitions and formats, and that can be readily accessed by users. Since the WCUD is still in early stages of development it currently incorporates only three distinct areas of reporting: (1) Enrolment, (2) Retention and Completion, and (3) University Income by Fund.

A benefit to Universities participating in the WCUD is the ability to measure success, or lack thereof, relative to peer institutions. Notwithstanding their value, however, the WCUD retention reports are of limited value in terms of meeting the numerous internal demands for retention and completion data. In light of these factors, ORP has sought to develop a reporting tool to meet both WCUD reporting requirements as well as internal requests. The result is a Retention Report Generator, using MS Access that allows users the option of producing one of two types of retention data reports (in Excel format) that can be used for retention and completion analysis.

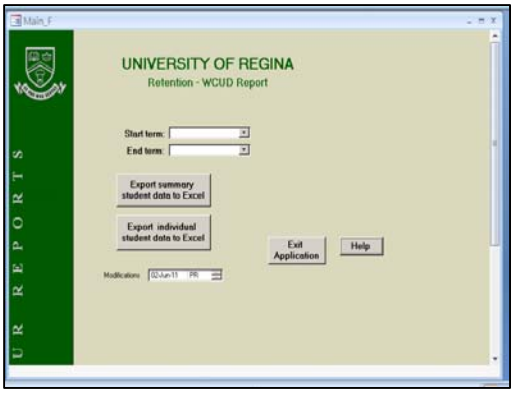


Figure 6 - Retention Report Main Page

At the EAIR presentation, source files produced from the Access Retention Report Generator will be used to demonstrate how Excel and Tableau pivot tables, and other features, can be used to explore various dimensions of student retention and completion among University of Regina students.

For reference purposes, a description of the two reports is provided below, along with a list and description of the data fields contained in the WCUD Summary Report Source file. It should be noted that the detailed report is still under development, and the existing beta version is currently lacking many fields needed for a more comprehensive assessment.

WCUD Summary Report:

- The WCUD report provides retention and completion rates of first-year, full-time, degree-seeking students entering the institution in a Fall term (or within a specified range of Fall terms) by program of study, gender and international status.

- It generates an Excel file containing 16 distinct fields showing student counts, retention (year 2, year 3) and completion (in 4, 5, 6, and 7+ years) by cohort year, credential, CIP code, gender, and international status.
- Each row in this report represents a cohort of students based on having a common credential, gender and International student status. A benefit of a summarized report, like this one, is the reduced number of records it creates. A shortcoming is less flexibility for exploring the data.

Detailed Student Retention Report:

- The detailed student report also provides retention and completion rates of first-year, full-time, degree-seeking students who enter the institution in a Fall term, or within a specified range of Fall terms. It generates many of the same fields as the WCUD Summary Report, as well as 10 new fields.
- It differs from the WCUD report in that each row represents a unique student, as opposed to a cohort of students who share common characteristics. Because each row represents a student, rather than a cohort, it contains a substantially larger number of records. This better enables multi-dimensional and detailed analysis, including data quality analysis aimed at vetting data files for completeness and accuracy.

Fields included in the WCUD summary report:

- ✓ **Cohort Year:** The entering year of students starting in a Fall term.
- ✓ **Institution** (Abbreviation): An acronym of up to four characters representing the institution.
- ✓ **Institution Name:** The full name of the institution
- ✓ **Gender:** The gender of a student as recorded by the institution, where F=Female, M=Male and U=Unknown
- ✓ **International (or Intl_R):** Distinguishes between students who have Canadian citizenship or permanent residence status in Canada (*non-international* students; expressed as "N") and students who have been authorized for study in Canada by holding a valid Study Permit or Work Permit (*Internationa*l students; expressed as "Y").
- ✓ **CIP:** A six-digit code in the form xx.xxxx that identifies instructional programs within educational institutions.
- ✓ **Credential:** The credential declared by a degree-seeking student when first registering (e.g. Bachelor of Arts, English); at the UR this is based on a combination of a student's declared *Program* and **Major**.
- ✓ **Cohort_Count** (Fall Cohort): A group of students entering in a given fall term. For WCUD reporting this includes all undergraduate full-time, first-time, degree-seeking students with a High School basis of admission; counts are by cohort year, credential, gender and international status.
- ✓ **Retained _YR2:** Count of students from the Entry cohort that were registered as of Nov. 15th in the following Fall term (Year Two).
- ✓ **Retained_YR3:** Count of students from the Entry cohort that were registered as of Nov. 15th in the third Fall term (Year Three)
- ✓ **Graduated_YR4:** Count of students from the Entry cohort that graduated with an undergraduate degree within four years (e.g. if the entry cohort year is Fall 2005, then the four year graduation count would include students who graduated by or in 2009).
- ✓ **Graduated_YR5:** as above, within five years.
- ✓ **Graduated_YR6:** as above, within six years.
- ✓ **Graduated_YR7_PL:** Count of students from the Entry cohort that graduated with an undergraduate degree within seven or more years.
- ✓ Institution Province: Two character province abbreviation (e.g., SK for Saskatchewan)

PivotTable 101: Basic Concepts and Terminology

A pivot table does not store its own data, it requires a **data source**. This fact can be obscured by the fact that a single Excel spreadsheet can contain both the data source and the pivot table, but they are physically and logically distinct, typically at least each in their own worksheet. As will be discussed in the presentation, Excel is most typically used for calculations or a series of linked calculations (i.e. a 'model'). Such calculations can feed into a pivot table, or use the results of a pivot table, or both. But again, the calculations are logically and physically distinct from the pivot table (although some very basic calculations can be performed *within* pivot tables). And finally, Excel can produce graphs. Again, these often use the results of a pivot table, but are logically and physically distinct from it (with the exception of the very limited 'pivot graphs').

Tableau is purely a pivot table. It does not store its own data. However, it can connect to data stored in Excel, and this is the method we will use in our example. It can also connect to a wide range of database systems (as can Excel, although typically not well). Unlike Excel, graphical display of data is completely integrated into Tableau. As will be illustrated in our presentation, it can be dragged and dropped and flipped and otherwise played with just like a pure numeric display. But Tableau does not do calculations (again with the exception of some very basic calculations *within* the pivot tables). In our example, we bring Tableau results into Excel for further calculation.

Data Source:

- The source data is the raw data underlying a PivotTable. The first step in creating a PivotTable is to organize the source table in a list format consisting of ***rows*** and ***columns***. The first row contains ***field*** headings. A field is a category or type of data. Rows contain data elements pertaining to cases. Each row represents a unique case (e.g., a student, or a cohort of students). Each case should have a unique identifier.
- Item – In a pivot table, an ***item*** is a subcategory of a field. Each item represents a unique value found in the field in the source data (see examples in Figure 4).

	A	B	C	D	E	F	G
1	Term	Unique ID	Gender	Time_ status	Faculty	Faculty2	Program
2	200030	123456781	F	FT	AR	Arts	ARUND
3	200030	123456782	F	FT	AR	Arts	ARBA
4	200030	123456783	F	FT	AR	Arts	ARUND
5	200030	123456784	F	FT	AR	Arts	ARUND
6	200030	123456785	F	FT	AR	Arts	ARPREPROF
7	200030	123456786	M	FT	AR	Arts	ARUND
8	200030	123456787	F	FT	ED	Education	EDELEM
9	200030	123456788	F	FT	SW	Social Work	SWQUAL
10	200030	123456789	M	FT	SC	Science	SCUND
11	200030	123456790	F	FT	AR	Arts	ARBA
12	200030	123456791	F	FT	FA	Fine Arts	FABA
13	200030	123456792	F	FT	FA	Fine Arts	FABMUS
14	200030	123456793	F	FT	ED	Education	EDELEM

Figure 1 - Source data from an Excel worksheet

In this example, the source data contains demographic and program information about a group of students. Each row represents a unique student and provides details about the student. These details (also known as ***items*** or **values**) can be expressed in numeric or text format. A good practice in preparing for Pivot Table analysis is to replace coded values with actual text values, or to create an extra field with the text data.

- The data source must be 'connected' to the pivot table. We will demonstrate this in the presentation. Note that source data in Excel is called a 'list', which must not contain any blank rows or columns (a very common cause of connection problems).

PivotTable Layout:

- Once the source data is prepared and connected, the next step in creating a pivot table is to decide what data (or fields) you want to use in the analysis. Whether using Excel or Tableau the basic data layout concept is the same even though the visual layouts differ, and some of the terminology and functionality are different.

- In Excel the ***drop area*** (outlined in blue) shows where fields from the field list window are placed. The ***field list,*** shown on the right of the screen, lists all of the fields from the source data that can be used to produce the PivotTable report. The ***data area*** contains the summary data for the row and column fields. The ***page fields*** are used to filter the PivotTable to display one or more of the items in one or more page fields. The ***data fields*** (or data items) contain the values that are being measured (e.g., counted, summed, etc.). Typically, only data fields with numeric values are summed, averaged or counted. Data fields with text values *can* be counted, but it is more likely that such a field would be used to form a row or column, with the numbers of instances of each text value being counted by using the student's ID#, or similar field, as the data item (and using a 'count', not a sum).

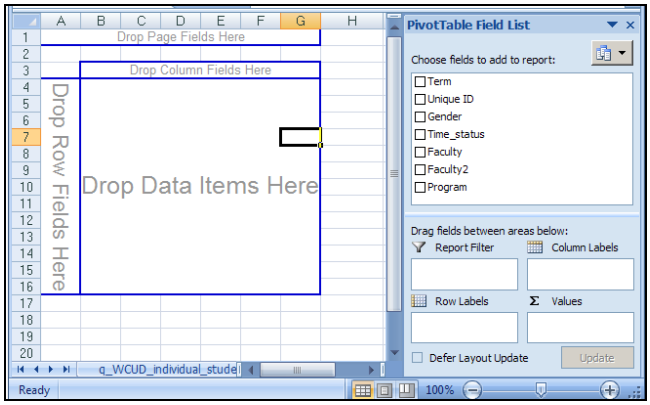


Figure 2 – Empty Excel Pivot Table

This empty Excel Pivot Table includes a field list on the right. These fields can be dragged and dropped directly into the **drop area** of the pivot table (which is outlined in blue) or into the drop boxes that appear below the fields list.

- In Tableau, fields from the source file are automatically grouped into either a **dimensions list** or a **measures list**, depending on whether the fields contain text or numeric values. Measures, which include fields with numeric values, are treated the same way as a data field in Excel.
- The ***filters shelf*** in Tableau operates in a similar fashion to the ***page area*** in Excel. This is where filters can be applied to a data dimension in order to narrow the range of items from a particular field, or set of fields, in a Tableau view.
- The ***pages shelf*** in Tableau is <u>not</u> used for filtering. It is used to spit a view into a sequence of pages based on the values in a field.

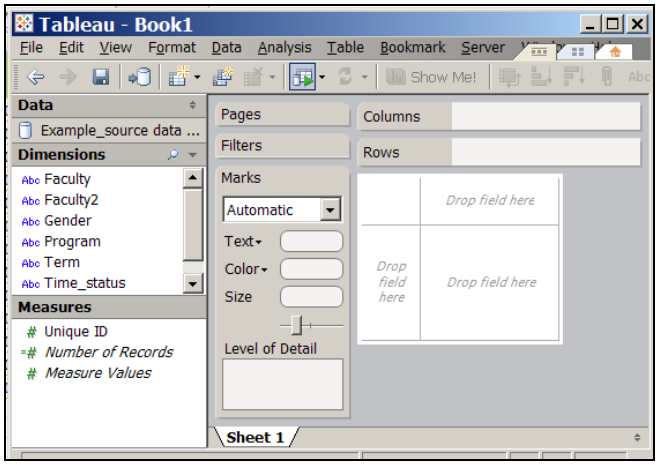
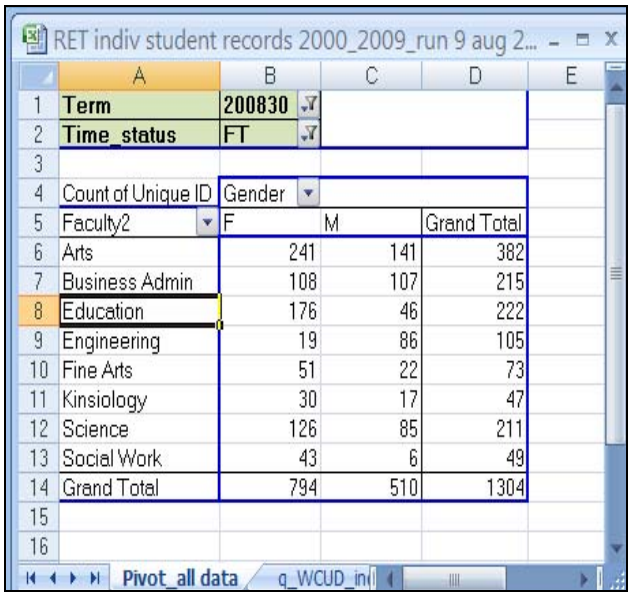


Figure 3 – Empty Tableau PivotTable

Tableau fields from the source file are listed on the left of the screen and are organized as either dimensions or measures. Dimensions are generally placed in a **column** or **row shelf**. Measures (or data fields) are typically dragged into the **text shelf**.



RET indiv student records 2000_2009_run 9 aug 2...

	A	B	C	D	E
1	Term	200830			
2	Time_status	FT			
3					
4	Count of Unique ID	Gender			
5	Faculty2	F	M	Grand Total	
6	Arts	241	141	382	
7	Business Admin	108	107	215	
8	Education	176	46	222	
9	Engineering	19	86	105	
10	Fine Arts	51	22	73	
11	Kinsiology	30	17	47	
12	Science	126	85	211	
13	Social Work	43	6	49	
14	Grand Total	794	510	1304	
15					
16					

Pivot_all data | q_WCUD_in

In this PivotTable **Term** and **Time Status** are **page fields**. Only items with the value "200830" (Fall 2008) and "FT" (Full-time) are summarized in the data.

The **data field** is a **unique student ID**. The column field is Gender and the row field is Faculty.

The **Gender items** include "F" and "M" and the **Faculty items** include the various sub-categories of "Arts", "Science," etc.

The cells in the **Data area** contain the summarized data. Using **Unique ID** as the (counted) **measure**, the table shows counts of the number of students by Gender and Faculty.

Figure 4 – Excel Pivot Table

- There is a **hierarchy** or **nesting** component to PivotTables that occurs when more than one field is added to a row or column. In the case of two or more row fields (see Figure 5), the field closest to the **data area** is called the **inner row** and all other row fields are called **outer rows**. Each type has different display properties. The items in the outermost row field (**Faculty2**) display only once, while the items in the innermost row field (**Term**) display as needed.

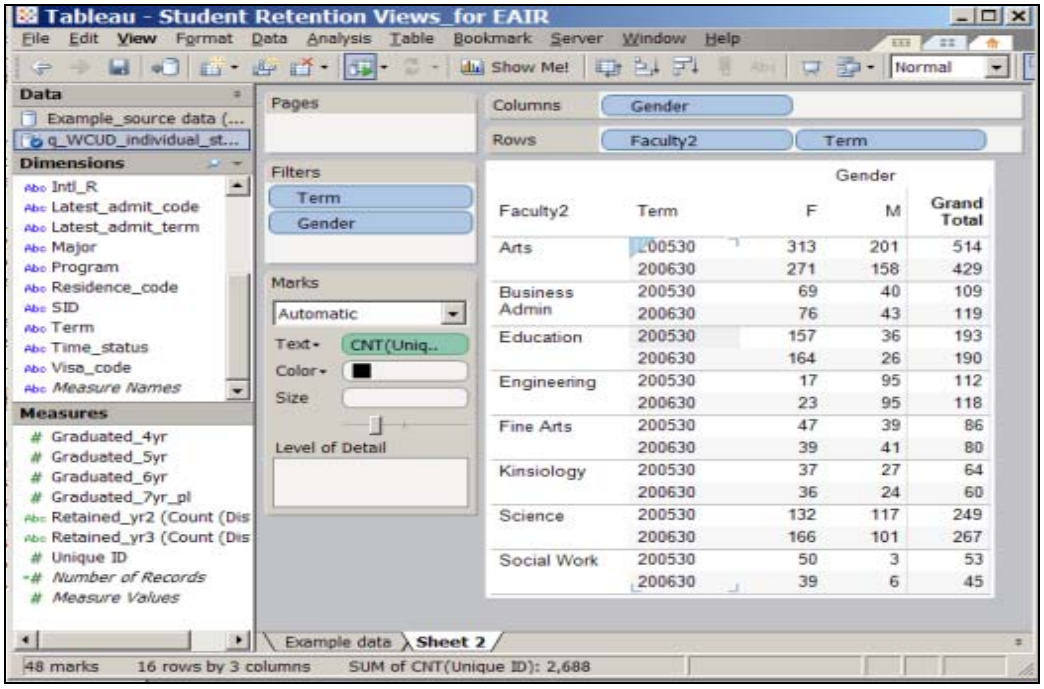


Figure 5 - Tableau Pivot with Two Row Fields

This brief overview has described the basic components of Excel and Tableau PivotTables. More terminology, concepts and tips will be introduced in the presentation. Our presentation and future discussion will also be posted to our website: **www.uregina.ca/orp/papers.shtml**