Creating Summary Tables

Rishi Valley 2018

	Women		Men	
	Number	*Avg. Rating*	*Number*	*Avg Rating*
All Movies	16,716	3.54	48,819	3.53
Star Wars	102	4.23	284	4.37
Abyss, The	20	4.00	82	3.55

Look at this table. It was created from the over 65,000 rows of data in the movie rating dataset we saw earlier. How long do you think it would take you to calculate the values in this table from the raw dataset of ~65,000 rows?

	Women		Men	
	Number	*Avg. Rating*	*Number*	*Avg Rating*
All Movies	16,716	3.54	48,819	3.53
Star Wars	102	4.23	284	4.37
Abyss, The	20	4.00	82	3.55

This is an example of a summary table. A lot of work and computation went into this.

Notice that this is actually new data that was computed from the raw data. This is way beyond filtering and sorting.

Computing this by hand for ~65,000 ratings (or writing formulas in a spreadsheet) would be pretty painstaking.

	Women		Men	
	Number	*Avg. Rating*	*Number*	*Avg Rating*
All Movies	16,716	3.54	48,819	3.53
Star Wars	102	4.23	284	4.37
Abyss, The	20	4.00	82	3.55

But we can use computing tools to create summary tables like this for us in a flash.

Most data manipulation tools, like spreadsheets, allow you to quickly group, categorize, count, and average things.

Making a summary table is a computational technique for exploring the data; let's try it.

Summary Tables

- A summary table typically represents one or more aggregations (groupings of items) and computations that are performed on the raw dataset.
- In most spreadsheet programs, a summary table is called a **pivot table.**
- Making a summary (pivot) table is often considered an advanced technique. Once you get used to it, however, it's an extremely powerful computational tool that is available in most spreadsheet software.

Vocabulary

- **Aggregation** - a computation in which rows from a data set are grouped together and used to compute a single value of more significant meaning or measurement. Common aggregations include: Average, Count, Sum, Max, Median, etc.
- **Pivot Table** - in most spreadsheet software it is the name of the tool used to create summary tables.
- **Summary Table** - a table that shows the results of aggregations performed on data from a larger data set, hence a "summary" of larger data. Spreadsheet software typically calls them "pivot tables".

Why?

- Being able to manipulate data is an important skill for computer scientists. Being able to create summary tables from larger datasets represents a form of computational thinking.
- To make a good summary table, one must have a good sense of the data, be able to hypothesize about what might be interesting to look at, and then have the skills to use a computational tool to create it.
- While seemingly mundane, a spreadsheet is an extremely powerful tool for working with data.
- Understanding the features of a spreadsheet tool, and what kinds of computations it can perform, can save you a lot of time and energy from either doing such things "by hand" or writing your own program to do it.

The Basics

Part 1 of Making Pivot Tables

Dataset

- Make a copy of Teenage Movie Ratings Subset
- Open it in a spreadsheet program
- The data is a subset of the larger movie ratings dataset we saw in a previous lesson.
- The dataset contains roughly 300 movie ratings that were collected online in 1997-98.
- The data has been filtered so that it only contains movies that were rated by at least 2 females and 2 males in the 14-18 year-old range.

Making Summary Tables

- We're first going to make a simple summary table that shows the average rating for every movie that's in the data.
- Here is what we're going for:

Movie	Avg. Rating
Air Force One	3.615384615
Chasing Amy	4
Contact	4.05
...	...
Toy Story	3.615384615
Twelve Monkey	4
Willy Wonka	3.818181818

Your first Summary table

- In most spreadsheet programs a Summary Table is called a pivot table.
- Do This: choose Data (or Insert) -> Pivot table…
- Add Rows and Values to Your Table
- The menu on the right side of the pivot table lets you choose what you want the rows, columns, and values to be in your summary table. We want to set it up so that:
- Each row is one movie.
- Each value is the average rating of that movie.

Visual Check

3	Row Labels	Average of rating
4	Air Force One	3.62
5	Chasing Amy	4.00
6	Contact	4.05
7	Courage Under Fire	3.64
8	E.T. the Extra-Terrestrial	3.80
9	Evita	3.36
10	Fargo	4.42
11	Game, The	4.33
12	Independence Day (ID4)	3.93
13	Liar Liar	3.13
14	Mission: Impossible	3.92
15	Phenomenon	3.54
16	Return of the Jedi	4.61
17	Rock, The	4.18
18	Saint, The	3.55
19	Scream	4.15
20	Star Wars	4.73
21	Titanic	4.67
22	Toy Story	3.62
23	Twelve Monkeys	4.00
24	Willy Wonka and the Chocolate Factory	3.82
25	**Grand Total**	**4.00**

What happened?

- **Computation!**
- The power of the pivot table is that it allows you to compute things you could never do by just filtering and sorting. The pivot table is doing a lot of computing behind the scenes for you - which is great - but you should understand what's really happening so you can make your own choices in the future. Here's a synopsis:
- Rows - Group By: movie
  - Rows act like the major categories or groupings for which you want to calculate values.
  - The Computation: When you set the rows to be "movie," the software finds all of the unique movie titles in the raw dataset and puts one on each row. This is called aggregation, which is a fancy word that means grouping or clustering.
- Values - Display: rating; Summarize by: AVERAGE
  - Values lets you specify the computation that should happen for each row.
  - The Computation: We're interested in the average rating for each movie, so for Values we choose rating, Summarize by: AVERAGE.

Lets change

- Let's Change the Value - Summarize by: COUNT
- Change summarize by from AVERAGE to COUNT.
- Now, instead of computing the average rating, this will count the number of ratings for each movie.

Visual Check

3	Row Labels	Count of rating
4	Air Force One	13.00
5	Chasing Amy	13.00
6	Contact	20.00
7	Courage Under Fire	11.00
8	E.T. the Extra-Terrestrial	10.00
9	Evita	11.00
10	Fargo	12.00
11	Game, The	12.00
12	Independence Day (ID4)	14.00
13	Liar Liar	16.00
14	Mission: Impossible	12.00
15	Phenomenon	13.00
16	Return of the Jedi	18.00
17	Rock, The	11.00
18	Saint, The	11.00
19	Scream	26.00
20	Star Wars	22.00
21	Titanic	12.00
22	Toy Story	13.00
23	Twelve Monkeys	11.00
24	Willy Wonka and the Chocolate Factory	11.00
25	**Grand Total**	**292.00**

Add another field!

- Add Another Field to Values
- Let's show both the average rating and the count side-by-side in the table. To do this we add another Values field.
- The count is already there, so let's add the average rating again.
- Now, for each movie we'll see the total number of ratings the average rating.

Visual Check

3	Row Labels	Count of user id	Average of rating
4	Air Force One	13.00	3.62
5	Chasing Amy	13.00	4.00
6	Contact	20.00	4.05
7	Courage Under Fire	11.00	3.64
8	E.T. the Extra-Terrestrial	10.00	3.80
9	Evita	11.00	3.36
10	Fargo	12.00	4.42
11	Game, The	12.00	4.33
12	Independence Day (ID4)	14.00	3.93
13	Liar Liar	16.00	3.13
14	Mission: Impossible	12.00	3.92
15	Phenomenon	13.00	3.54
16	Return of the Jedi	18.00	4.61
17	Rock, The	11.00	4.18
18	Saint, The	11.00	3.55
19	Scream	26.00	4.15
20	Star Wars	22.00	4.73
21	Titanic	12.00	4.67
22	Toy Story	13.00	3.62
23	Twelve Monkeys	11.00	4.00
24	Willy Wonka and the Chocolate Factory	11.00	3.82
25	**Grand Total**	**292.00**	**4.00**

That's It for the Basics of Pivot Tables!

- There's not much more to it than that. Once you get the hang of pivot tables they can be a very powerful tool for manipulating data. There are more advanced things you can do with a pivot table if you like, but you know enough now that you can probably just play around with the other settings and see what happens. **Key Ideas:**
- **Summary tables (pivot tables) provide a way to visualize data.** Yes, it's a table, but by aggregating and summarizing information from a large data set, summary tables allow you to see things in the data you might otherwise not see.
- **Summary tables allow you to manipulate and create new data**. Even for our simple movies example here, the raw data didn't contain the average rating for every movie, or count how many ratings there were. We had to compute it, and the pivot table let us do that quickly and easily.
- **A summary table helps you look at your data in new ways.** Think: how could data be grouped? What could be calculated? Once you know how to make a summary table you can begin to look at raw data and ask questions that you know might be possible to answer.
- **A summary table can be a first step toward a good visualization**. Often it's difficult to make a meaningful chart or graphic out of raw data. You often want to summarize it first, then chart it!

Manipulation and Visualization

Part 2 of Making Pivot Tables

Adding Columns

- Let's look at two more features of pivot tables that will allow you to do more complex investigations of your data.
- We learned that a Row in a pivot table specifies an aggregation or grouping of items for which you want to compute a value.
  A Column in a pivot table is just another aggregation, but it displays the values across the top of the table. It's easier to understand when you see it...
- Do This: Add columns that group your data by gender
- The resulting table shows the average rating and count for each movie, but also broken down by gender. The pivot table also preserves the "Grand Totals" which is what the data would look like if no columns were specified.

Visual Check

3		Column Labels					
4		F		M		Total Average of rating	Total Count of user id
5	Row Labels	Average of rating	Count of user id	Average of rating	Count of user id		
6	Air Force One	3.20	5.00	3.88	8.00	3.62	13.00
7	Chasing Amy	4.33	3.00	3.90	10.00	4.00	13.00
8	Contact	3.43	7.00	4.38	13.00	4.05	20.00
9	Courage Under Fire	3.80	5.00	3.50	6.00	3.64	11.00
10	E.T. the Extra-Terrestrial	4.00	4.00	3.67	6.00	3.80	10.00
11	Evita	3.50	6.00	3.20	5.00	3.36	11.00
12	Fargo	4.00	2.00	4.50	10.00	4.42	12.00
13	Game, The	3.80	5.00	4.71	7.00	4.33	12.00
14	Independence Day (ID4)	4.57	7.00	3.29	7.00	3.93	14.00
15	Liar Liar	3.14	7.00	3.11	9.00	3.13	16.00
16	Mission: Impossible	4.33	3.00	3.78	9.00	3.92	12.00
17	Phenomenon	3.50	6.00	3.57	7.00	3.54	13.00
18	Return of the Jedi	4.67	3.00	4.60	15.00	4.61	18.00
19	Rock, The	4.00	4.00	4.29	7.00	4.18	11.00
20	Saint, The	4.20	5.00	3.00	6.00	3.55	11.00
21	Scream	4.40	10.00	4.00	16.00	4.15	26.00
22	Star Wars	4.83	6.00	4.69	16.00	4.73	22.00
23	Titanic	4.80	5.00	4.57	7.00	4.67	12.00
24	Toy Story	4.60	5.00	3.00	8.00	3.62	13.00
25	Twelve Monkeys	2.50	2.00	4.33	9.00	4.00	11.00
26	Willy Wonka and the Chocolate Factory	4.00	4.00	3.71	7.00	3.82	11.00
27	**Grand Total**	**4.00**	**104.00**	**3.99**	**188.00**	**4.00**	**292.00**

Filtering Pivot Tables

- Applying a filter to a pivot table does the same thing as it does in the normal spreadsheet - it allows you to filter out values from the raw data.
- Do now: first filtering out 14-year-olds from the calculations, and then filter out some of the movies.
- You don't have to do this, but in some instances it can be a very useful tool.

Visual Check

2	age	(Multiple Items)					
3							
4		Column Labels					
5		F		M		Total Average of rating	Total Count of user id
6	Row Labels	Average of rating	Count of user id	Average of rating	Count of user id		
7	Rock, The	4.00	4.00	4.29	7.00	4.18	11.00
8	Titanic	4.75	4.00	4.57	7.00	4.64	11.00
9	**Grand Total**	**4.38**	**8.00**	**4.43**	**14.00**	**4.41**	**22.00**
10							

Manipulating the Pivot Table

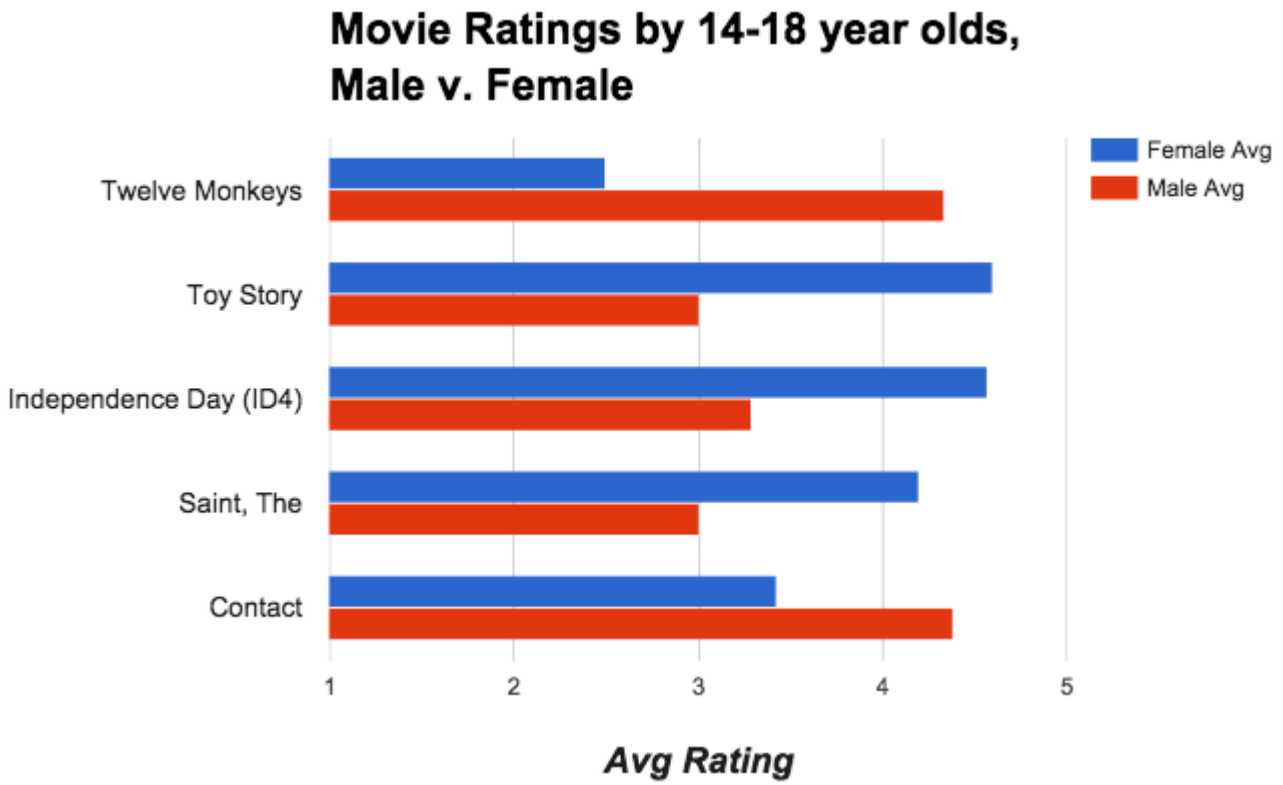
- If you want to maninpulate the data further, to sort or filter, you shouldn't do it in the live, active pivot table. Instead you should copy the table, and paste the values into a new spreadsheet.
- Note: "Paste Values" is not the same as a normal "Paste".
- Do This: Copy the pivot table, create a new tab in the spreadsheet and do Edit -> Paste special -> Paste values only.
- "Why Paste values only instead of just Paste?"
- If you copy a pivot table and do a normal paste it will paste another copy of the active, live, responsive pivot table into a new tab. We don't want the active table; we just want the values it produced.
- You probably want to add/change column headings to display the table, especially to use it for charting.

Visual Check

	A	B	C	D	E	F	G
1	Movie	Avg Female Ratin	Female Viewers	Avg Male Rating	Male Viewers	Average Rating	Total Viewers
2	Air Force One	3.20	5	3.88	8	3.62	13
3	Chasing Amy	4.33	3	3.90	10	4.00	13
4	Contact	3.43	7	4.38	13	4.05	20
5	Courage Under Fire	3.80	5	3.50	6	3.64	11
6	E.T. the Extra-Terrestrial	4.00	4	3.67	6	3.80	10
7	Evita	3.50	6	3.20	5	3.36	11
8	Fargo	4.00	2	4.50	10	4.42	12
9	Game, The	3.80	5	4.71	7	4.33	12
10	Independence Day (ID4)	4.57	7	3.29	7	3.93	14
11	Liar Liar	3.14	7	3.11	9	3.13	16
12	Mission: Impossible	4.33	3	3.78	9	3.92	12
13	Phenomenon	3.50	6	3.57	7	3.54	13
14	Return of the Jedi	4.67	3	4.60	15	4.61	18
15	Rock, The	4.00	4	4.29	7	4.18	11
16	Saint, The	4.20	5	3.00	6	3.55	11
17	Scream	4.40	10	4.00	16	4.15	26
18	Star Wars	4.83	6	4.69	16	4.73	22
19	Titanic	4.80	5	4.57	7	4.67	12
20	Toy Story	4.60	5	3.00	8	3.62	13
21	Twelve Monkeys	2.50	2	4.33	9	4.00	11
22	Willy Wonka and the Chocolate Factory	4.00	4	3.71	7	3.82	11
23	Grand Total	4.00	104	3.99	188	4.00	292

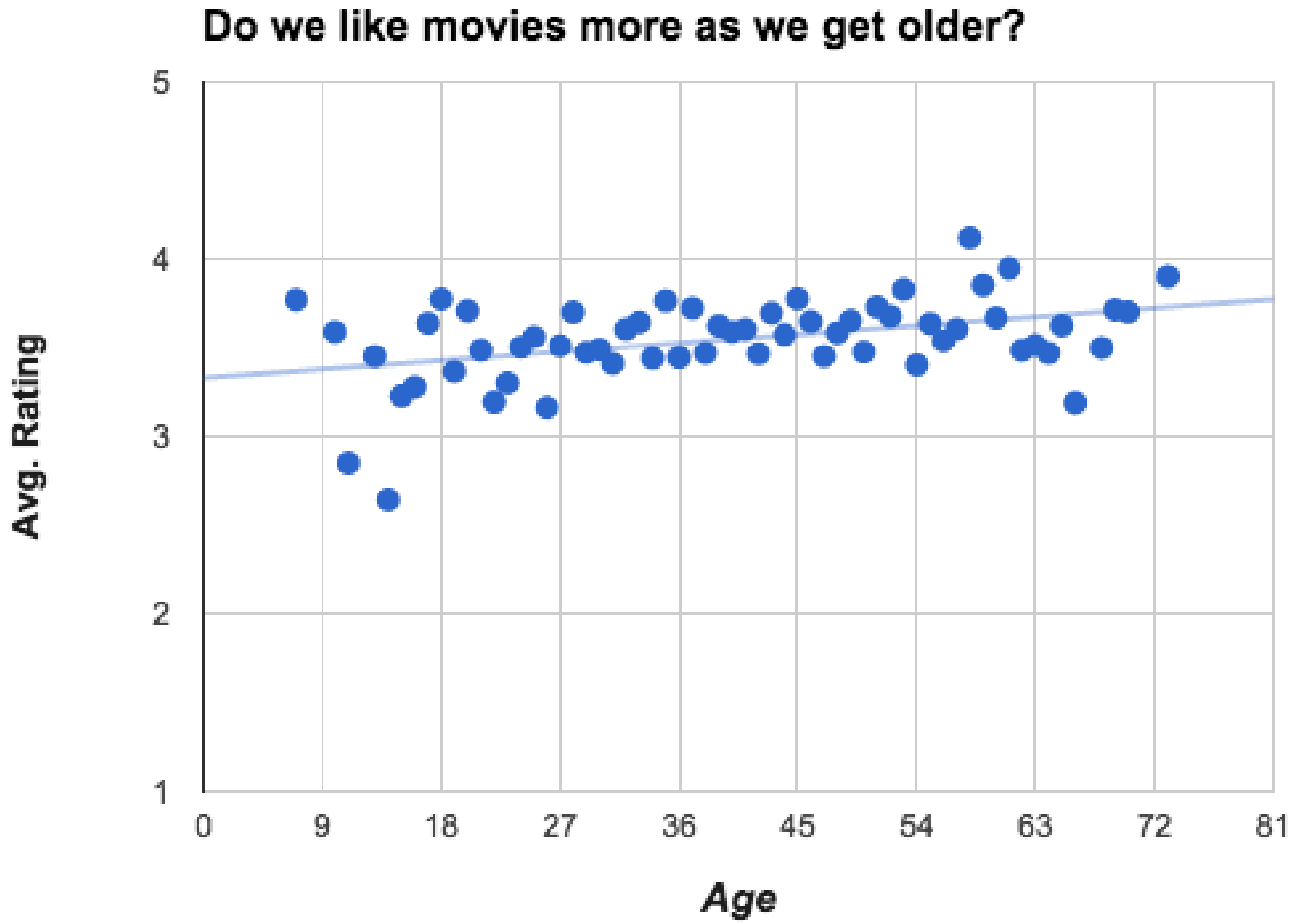
Make a Chart

From cleaning that up, plus some filtering we can make a chart of movies where the differences between male and female ratings are significant.



Moving on: Visualizing Summary Tables

- A summary table can be a good first step toward a great visualization. You often want to summarize data first, then chart it, so you can see larger connections or patterns.
- Summary tables also don't have to be small! You might make a summary table that is still too big or full of numbers to see any trends in the data.
- For example, from the original movie rating data (which had roughly 65,000 records) if you make a pivot table that shows the average movie rating for every possible age group the table will be about 75 rows long with a whole bunch of decimal numbers.
- You can't see any trend or pattern in the data just by looking at the table. But if you plot the results on a graph you can!



NOTE: A deeper investigation of the data shows that the number of movies rated by people at this web site declined steadily after age 28. The upward trend may be affected by the fewer number of ratings.

Free Play!

Now you'll make summary tables of the data you collected and cleaned!

- **Task:** Create at least two (2) pivot tables that show different things about your data. With your partner, go back to the data you collected as a class (and which you cleaned up yesterday). Practice using pivot tables to group and calculate things you might be interested in.
- **Tips:** There are two approaches to thinking about what kind of summary table to make:
- **Work backward from a question:** Start with a question you want to answer, or a hypothesis about something in the data you think you could reveal. Often the question itself tells you what calculations you need to make.
- **Work forward by experimenting, iterating, and finding something interesting.** Start by simply picking a category to group in rows. Then pick a second one to display as values, and try COUNT, AVERAGE, MIN, MAX, etc. By poking around ideas will come to you for interesting investigations.

Summary

Key ideas of Summary Tables

- **Summary tables (pivot tables) provide a way to visualize data.** Yes, it's still a table, but by aggregating and summarizing information from a large dataset, summary tables allow you to see things in the data you might otherwise not see.
- **Summary tables allow you to manipulate and create new data.** Even for our simple movies example here, the raw data didn't contain the average rating for every movie, or count how many ratings there were. We had to compute it, and the pivot table let us do that quickly and easily.
- **A summary table helps you look at your data in new ways.** Think: how could data be grouped? What could be calculated? Once you know how to make a summary table you can begin to look at raw data and ask questions that you know might be possible to answer.
- **A summary table can be a first step toward a good visualization** Often it's difficult to make a meaningful chart or graphic out of raw data. You often want to summarize it first, then chart it!

Quiz

- **Which of the following statements are true about pivot tables?**
- **Select two answers.**
- Pivot tables are used to quickly remove errors and inconsistencies from a dataset.
- Pivot tables are used to quickly perform aggregate computations and groupings on a set of raw data
- Pivot tables are used because they automatically detect and highlight potential trends or patterns in the underlying raw data
- Pivot tables are used to generate a summarized view of a large dataset which is helpful for gaining insight