Protobi™ User Guide

info@protobi.com

Table of contents

Welcome

You've fielded a survey and now want to see the data. To see it yourself, find the story, clean the data, and write an insightful analysis. To share it with your client, who really wants to see the data they bet a good part of their research budget to collect. To answer followup questions and spark new insights.

Protobi is a useful tool to see and explore survey data. To see data as distributions, and run comparisons, contrasts and crosstabs. It's specifically designed for surveys and handles a lot of the practical routine cases.

As much as we've designed the tool for you, it's really the tool that we wanted for ourselves, to fit the space between spreadsheets on one hand and statistical packages on the other.

Once you get familiar with it, it's a totally new way of seeing the data and a sharp tool in your hands. You'll find new "Aha!" moments in the data, catch more outliers that might have slipped through, and have a lot more fun with the followup questions that good answers inevitably generate.

This tutorial shows you everything you need to know to use Protobi, as a viewer, a designer or an administrator.

Sign in to Protobi

To sign in via email, enter your email address and click "Sign in via email". You'll receive an email with a link that will sign you in, and keep you signed in for 24 hours (or until you sign out).

It's like the "Forgot my password" link in many other websites, except that we never require you to remember one. This helps us keep your account information secure.

Sign in or create a new account

email

✉ Email

g+ Google

in LinkedIn

By signing in, you agree to the Terms of Use.
Why this is safe...

Alternatively, use Open ID for single-click sign-in. If your email is a Google email (i.e. either Google Apps for Business or Gmail) or if your email account is associated with your Linked In account, then these offer secure and convenient ways to sign in.

Open a project

When you login to Protobi, you'll see a list of your projects, and a list of projects that others have shared with you. ("Your" projects are those where you have admin privileges, and "Shared" projects are those where you have view/edit privileges).

A few public projects are automatically shared with you when you first login to serve as examples. We'll be using the Gender and Generations Survey from Pew Research as the basis for examples in this tutorial.
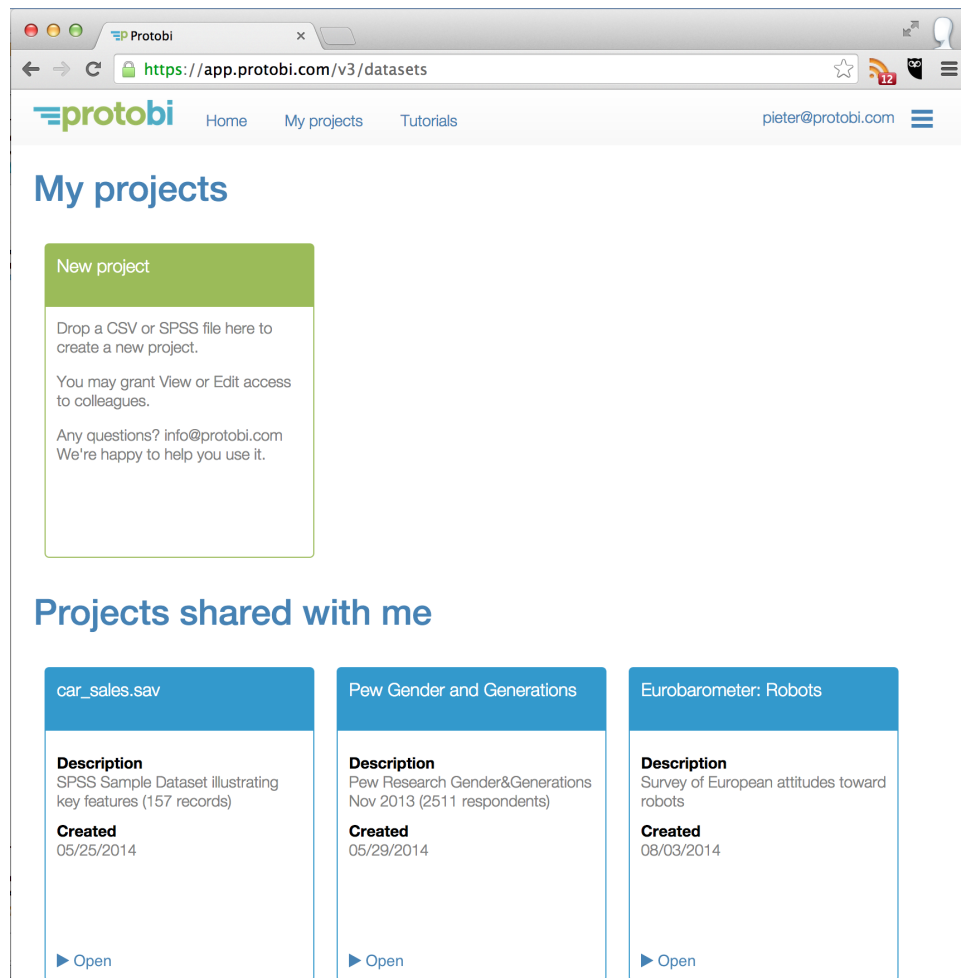
The Protobi interface

The Protobi interface has three main areas:

- The **Panel** in the center displays data graphically as interactive charts.
- The **Tree** on the left organizes data conceptually into a tree.
- The **Toolbar** on the top shows the current sample and controls how data is displayed.
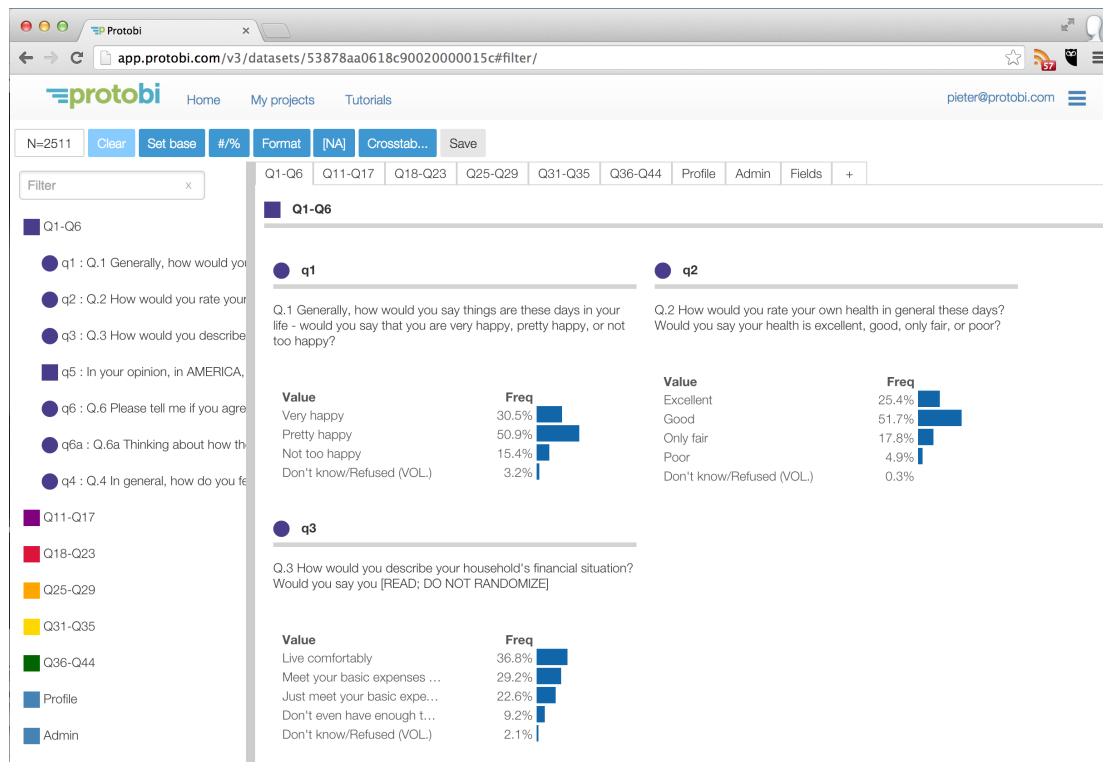
Tree

The Tree organizes data hierarchically into conceptual groups. The circle icons ● represent individual fields or "elements". The square icons ■ represent collections of elements or "groups".

In the example here, the first group "Q1 - Q6" contains the first ten questions in the survey, and is shown expanded. Most of these are individual elements, shown with circle icons. But "Q5" is itself a collection of subquestions and is shown as a square icon.

Note that these groupings are driven by the survey content. Another survey may have groups for "S. Screener", "A. Demographics" and so on.

These icons are clickable. In the Tree, click on the icons to expand/collapse groups, or right-click to bring up the context menu for editing options.

You can minimize the Tree by clicking on the border between the Tree and the Panel.

Use the search feature to quickly find elements by text. For instance, you can find any question that has the word "happy" or "work" in the title.

Panel

The Panel displays the data graphically as interactive charts. This is the main interface designed to see and explore the data.
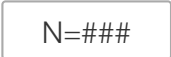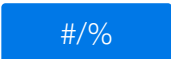
The panels are organized into tabs. Each top-level group in the Tree is also a tab in the Panel. Typically these will be logical associations by content, like sections of a survey.

Toolbar

The Toolbar has a number of buttons to help navigate the data:

N=2511	Clear	Set base	#/%	Format	[NA]	Crosstab...	Save
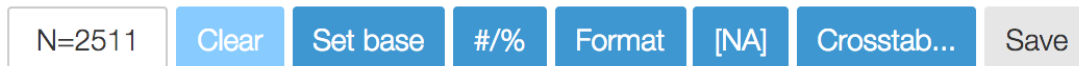
Hover over any of the toolbars to see its function and keyboard shortcut.

- **N=###**   shows the current sample size. Click to see filters in the current scenario.
- **Clear**   clears all filters, setting the current scenario to all respondents .
- **Set base**   sets baseline scenario for comparisons to the current scenario.
- **#/%**   toggles display between percentages and counts.
- **Format**   toggles display between formatted and raw data values.
- **[NA]**   toggles whether blank/missing/NA values are included in percentages.
- **Save**   saves any changes [requires EDIT permissions]

Explore the data

Click any value to filter data

Explore the data by clicking on any value to subset the data and see visual comparisons.

Here are elements showing values for the first two questions in the Gender and Generations survey. Here we can see that 30.5% of respondents rate themselves as "Very happy", and 25.4% rate themselves as being in "Excellent" health:

● **q1**

Q.1 Generally, how would you say things are these days in your life - would you say that you are very happy, pretty happy, or not too happy?

Value	Freq
Very happy	30.5%
Pretty happy	50.9%
Not too happy	15.4%
Don't know/Refused (VOL.)	3.2%

● **q2**

Q.2 How would you rate your own health in general these days? Would you say your health is excellent, good, only fair, or poor?

Value	Freq
Excellent	25.4%
Good	51.7%
Only fair	17.8%
Poor	4.9%
Don't know/Refused (VOL.)	0.3%

Below shows results for **q1** and **q2** after clicking on "Excellent" under **q2**:

● **q1**

Q.1 Generally, how would you say things are these days in your life - would you say that you are very happy, pretty happy, or not too happy?

Value	Freq
Very happy	49.6%
Pretty happy	42.7%
Not too happy	6.0%
Don't know/Refused (VOL.)	1.7%

● **q2**

Q.2 How would you rate your own health in general these days? Would you say your health is excellent, good, only fair, or poor?

Value	Freq
Excellent	100.0%
Good	0.0%
Only fair	0.0%
Poor	0.0%
Don't know/Refused (VOL.)	0.0%

In this view, we're drilled into a subset of respondents, 100% of whom selected "Excellent" to **q2**. This is highlighted in gold because this is an active filter.

When you create a filter, all elements throughout the map update to reflect this subset. For instance, we can see that 49.6% of these respondents selected "Very happy" to **q1**. This is much higher than the overall response of 30.5% for all respondents.

The baseline distribution is shown as a light grey shadow, so you can easily see what's different. Baselines and statistical comparisons are discussed below.

Enter multiple filters

You can make multiple simultaneous filters. For instance, you can click on both "Excellent health" in **q2** and "Not too happy" in **q1** and see that these respondents are much less likely to have enough income to "Live comfortably" in **q3**:

● q1

Q.1 Generally, how would you say things are these days in your life - would you say that you are very happy, pretty happy, or not too happy?

Value	Freq
Very happy	0.0%
Pretty happy	0.0%
Not too happy	100.0%
Don't know/Refused (VOL.)	0.0%

● q2

Q.2 How would you rate your own health in general these days? Would you say your health is excellent, good, only fair, or poor?

Value	Freq
Excellent	100.0%
Good	0.0%
Only fair	0.0%
Poor	0.0%
Don't know/Refused (VOL.)	0.0%

● q3

Q.3 How would you describe your household's financial situation? Would you say you [READ; DO NOT RANDOMIZE]

Value	Freq
Live comfortably	15.8%
Meet your basic expenses …	28.9%
Just meet your basic expe…	34.2%
Don't even have enough t…	18.4%
Don't know/Refused (VOL.)	2.6%

See current and baseline filters

The "N=" button in the Toolbar shows the size of the currently selected samle turns green when there are active filters. Click the "N=" button in the Toolbar to enumerate all current filters:

N=38	Clear	Set base	#/%	Format	[NA]	Crosstab...

Current filters

q2	'Excellent'
q1	'Not too happy'

Baseline filters

(No baseline filters)
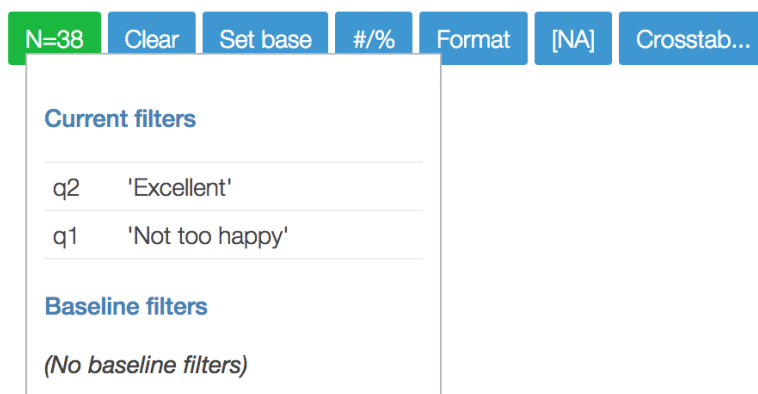
In the example here, we're filtering for those 38 respondents who answered both "Excellent health" in **q2** and "Not too happy" in *q1*

Advanced queries: 'Is', 'Or' and 'Not'

Click to select a value ('Is')

Ordinary clicks select a specific value, which we call 'Is' filters. For instance, here we've selected respondents where **q2** IS 'Excellent':

● **q2**

Q.2 How would you rate your own health in general these days?
Would you say your health is excellent, good, only fair, or poor?

Value	Freq
Excellent	100.0%
Good	0.0%
Only fair	0.0%
Poor	0.0%
Don't know/Refused (VOL.)	0.0%
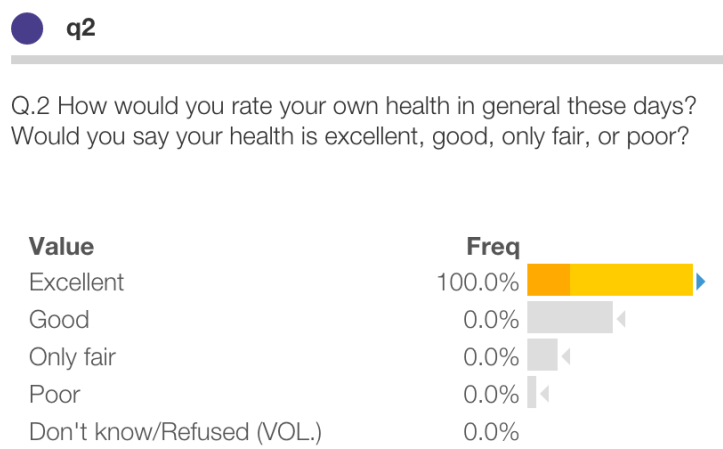
Shift+Click to exclude a value ('Not')

Hold the Shift key while clicking to do the opposite and exclude specific values, which we call 'Not' filters. For instance, here we've selected respondents where **q2** is NOT 'Excellent':

● **q2**

Q.2 How would you rate your own health in general these days?
Would you say your health is excellent, good, only fair, or poor?

Value	Freq
Excellent	0.0%
Good	69.3%
Only fair	23.8%
Poor	6.5%
Don't know/Refused (VOL.)	0.4%
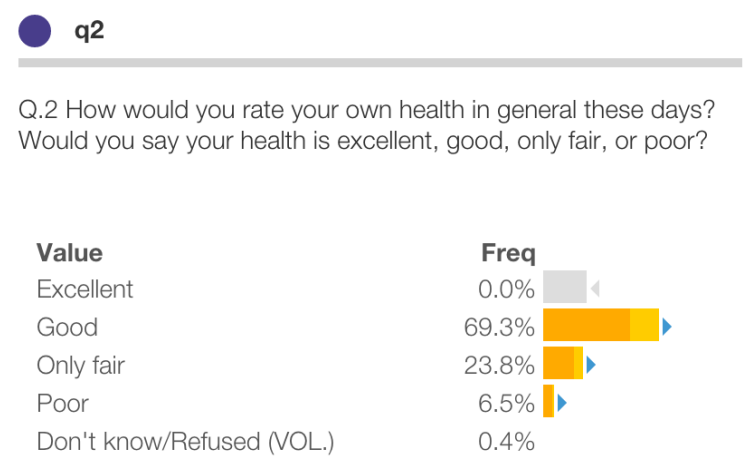
Option+Click to select multiple values ('Or')

Hold the Option (OSX) or Ctrl+Alt (Windows) key while clicking to select multiple values within an element, which we call 'Or' filters. For instance, here we've selected respondents where **q2** is 'Excellent' or 'Good':

● **q2**

Q.2 How would you rate your own health in general these days?
Would you say your health is excellent, good, only fair, or poor?

Value	Freq	
Excellent	32.9%	
Good	67.1%	
Only fair	0.0%	
Poor	0.0%	
Don't know/Refused (VOL.)	0.0%	
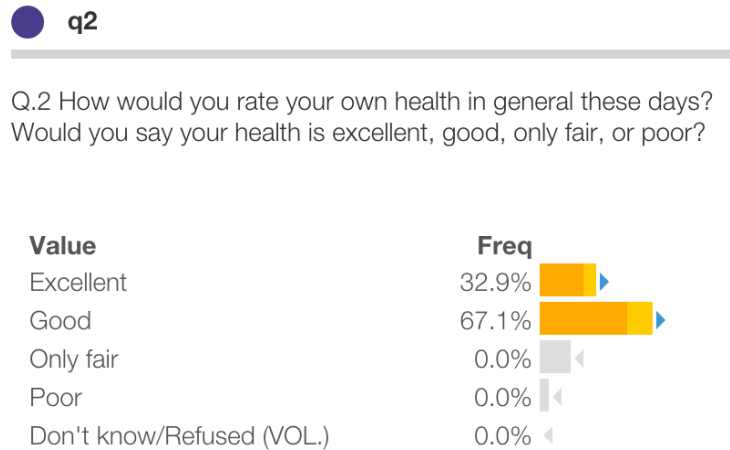
Clear filters

You can clear specific filters by clicking again on the value in the Panel. Or click the
Clear button in the Toolbar to clear all filters.

Baseline distribution

A powerful feature of Protobi is the ability to visually and statistically compare and contrast subsets of the data. You can set up one subset as a baseline, and compare another subset to it, and easily see the major differences.

The baseline distribution is shown as a light grey shadow for reference, to allow easy visual comparisons:

- Where the baseline and the current distributions overlap, they bars are dark blue.
- Where the current distribution exceeds the baseline, it is bright blue:
- Where the baseline distribution exceeds the current, it is light grey:

In the prior example, we can see that of respondents who rate themselves in "Excellent health" in **q2**, 49.6% rate themselves as "Very happy" in **q1**.

We can easily see that this is quite a bit higher than the baseline (which right now is all respondents) as there is quite a bit of bright blue extending past the baseline. We may further say that this difference appears statistically significant, as indicated by the blue triangle ▶.

q1

Q.1 Generally, how would you say things are these days in your life - would you say that you are very happy, pretty happy, or not too happy?

Value	Freq	
Very happy	49.6%	
Pretty happy	42.7%	
Not too happy	6.0%	
Don't know/Refused (VOL.)	1.7%	

q2

Q.2 How would you rate your own health in general these days? Would you say your health is excellent, good, only fair, or poor?

Value	Freq	
Excellent	100.0%	
Good	0.0%	
Only fair	0.0%	
Poor	0.0%	
Don't know/Refused (VOL.)	0.0%	

Statistical comparisons to baseline
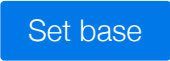
Triangle icons ◀▶ indicate statistically significant differences from baseline, at the 0.05 confidence level.

Thus, in this example, we may say that respondents who rate themselves in "Excellent health" in **q2** appear *significantly more* likely to rate themselves as "Very happy" in **q1**.

Similarly, these respondents appear *signficantly less* likely to rate themselves as "Not too happy."

If the current scenario a strict subset of the baseline scenario, Protobi intelligently subtracts the two populations to calculate a stronger significance test.

Set the baseline

This allows you to make strong contrasts and comparisons. A good practice is to ensure that the current and baseline distributions are disjoint (i.e. they don't overlap). You can set the baseline to the current filters at any time by pressing the button Set base .

For instance, we may wish to compare people whose health in **q2** IS "Excellent" versus those health is NOT "Excellent". This is a stronger contrast than in example at the start of this chapter, where the baseline includes *all* respondents, including some whose health is "Excellent".)

We can do that in a short sequence of steps:

1. Set current scenario to **q2** is NOT "Excellent"
2. Click "Set base" to make that our new baseline
3. Set current scenario to **q2** IS "Excellent"

We demonstrated step (1) in an earlier example by holding the Shift key while clicking on "Excellent" in q2. After clicking "Set base" in step (2) the distribution will look as below. Note that all the bars are dark blue and dark gold because the current scenario now equals the baseline scenario:

q1

Q.1 Generally, how would you say things are these days in your life - would you say that you are very happy, pretty happy, or not too happy?

Value	Freq
Very happy	24.0%
Pretty happy	53.6%
Not too happy	18.6%
Don't know/Refused (VOL.)	3.7%

q2

Q.2 How would you rate your own health in general these days? Would you say your health is excellent, good, only fair, or poor?

Value	Freq
Excellent	0.0%
Good	69.3%
Only fair	23.8%
Poor	6.5%
Don't know/Refused (VOL.)	0.4%

Finally, after clicking "Exellent" in step (3) the comparison will appear as follows.

q1

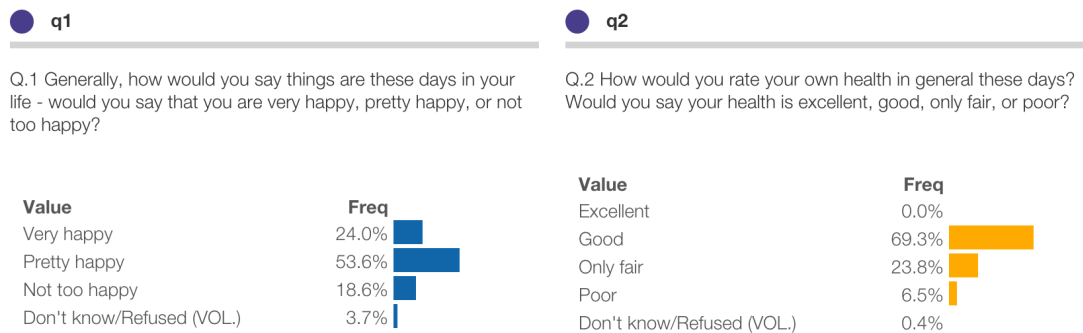Q.1 Generally, how would you say things are these days in your life - would you say that you are very happy, pretty happy, or not too happy?

Value	Freq
Very happy	49.6%
Pretty happy	42.7%
Not too happy	6.0%
Don't know/Refused (VOL.)	1.7%

q2

Q.2 How would you rate your own health in general these days? Would you say your health is excellent, good, only fair, or poor?

Value	Freq
Excellent	100.0%
Good	0.0%
Only fair	0.0%
Poor	0.0%
Don't know/Refused (VOL.)	0.0%

Note that the current distribution is the same as before (i.e. 49.6% of respondents in "Excellent" health are "Very happy") but the baseline distribution is different, and there are larger differences shown in bright blue or grey.

Further notice we get the same set of triangle icons indicating significant differences here as

before. This is because Protobi had intelligently subtracted the two populations to calculate a stronger significance test.

Don't overinterpret significance tests

Caution: Statistical significance is a famously tricky concept. As used here it is a narrow technical term, which means simply that this difference is large enough that it appears unlikely to happen by chance. So if this study were fielded with a different set of respondents from the same population, we'd likely see this difference again.

We keep using the word "appears" here because statistical hypothesis testing has strong limitations. Strictly, the test we're using here applies only to hypotheses you specify before looking at the data. Thus if you had hypothesized in advance that people in 'Excellent' health are more likely to be 'Very happy', this would be strong evidence in support of that hypothesis.

However, if you click to create a filter, and go scanning for "significant" differences, this test is less meaningful. By definition, we'd expect 5% of comparisons to appear as "significant". If we're scanning only for those differences that appear as "significant" after drilling in, and making up hypotheses the findings might support, many of these are likely to be false positives. It can still be instructive to do so, just not as strong evidence.

By analogy, consider the game of pool. If you call "9 ball into the side pocket off the bank" and then make a shot and that outcome happens, that's pretty impressive — it's likely you could do make that shot again in the same situation. On the other hand, if you just made the same shot without calling it ahead of time, that's not so impressive — it's not at all clear you could do that again.
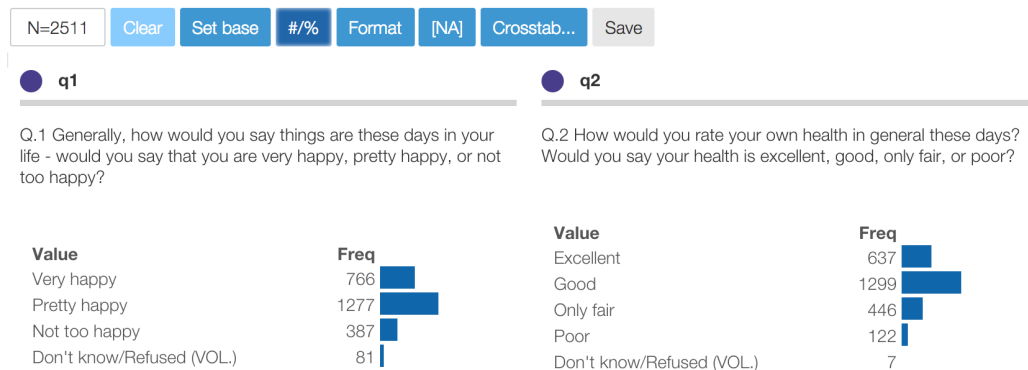
Further, just because a difference is statistically significant doesn't mean that you as a manager must make a decision to act. Even small numeric differences can be statistically significant with a large sample.

Net net: Use statistical significance as a qualitative way to quickly identify differences worth your attention.

Toggle display of data and frequencies

Show frequencies as counts

For example, click the **#/%** button to show frequencies as counts rather than percentages. Here we can see that 766 respondents answered "Very happy" to **q1**:

| N=2511 | Clear | Set base | #/% | Format | [NA] | Crosstab... | Save |

● q1

Q.1 Generally, how would you say things are these days in your life - would you say that you are very happy, pretty happy, or not too happy?

Value	Freq	
Very happy	766	
Pretty happy	1277	
Not too happy	387	
Don't know/Refused (VOL.)	81	

● q2

Q.2 How would you rate your own health in general these days? Would you say your health is excellent, good, only fair, or poor?

Value	Freq	
Excellent	637	
Good	1299	
Only fair	446	
Poor	122	
Don't know/Refused (VOL.)	7	

Show raw data values

Alternately, click the **Format** button to show the raw unformatted values rather than the value labels. In this example the raw data has values 1, 2, 3 and 9 for **q1** and the data file specifies value formats such as {"1":"Very happy", "2":"Pretty happy", ...}:

| N=2511 | Clear | Set base | #/% | Format | [NA] | Crosstab... | Save |

● q1

Q.1 Generally, how would you say things are these days in your life - would you say that you are very happy, pretty happy, or not too happy?

Value	Freq	
1	30.5%	
2	50.9%	
3	15.4%	
9	3.2%	

● q2

Q.2 How would you rate your own health in general these days? Would you say your health is excellent, good, only fair, or poor?

Value	Freq	
1	25.4%	
2	51.7%	
3	17.8%	
4	4.9%	
9	0.3%	
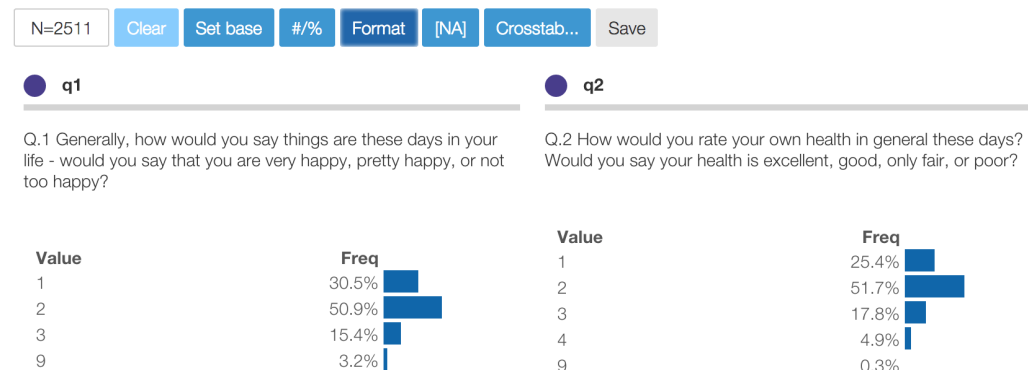
Blank/missing/skipped values: [NA]

NA is a special value which represents "Not available", "Not applicable", or "No answer". It can occur when:

- Someone skips a question due to a skip pattern
- Survey logic doesn't force an answer
- A profile variable has no recorded value
- The analyst removes outliers for individual questions

Sometimes we want to explicitly consider [NA] as meaningful values, other times we may wish to exclude them from analysis.

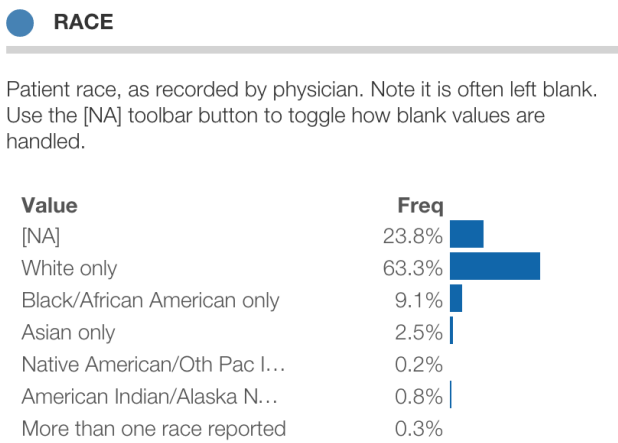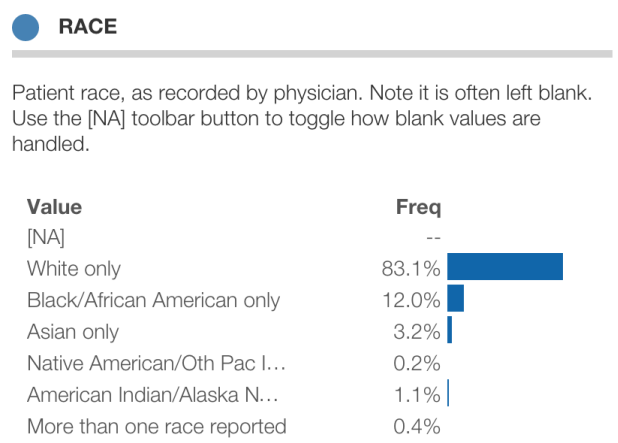Below is the distribution of patient race as recorded by US physicians in outpatient visits. Here can see that race is not recorded for 23.8% of patient visits:

● **RACE**

Patient race, as recorded by physician. Note it is often left blank.
Use the [NA] toolbar button to toggle how blank values are handled.

Value	Freq
[NA]	23.8%
White only	63.3%
Black/African American only	9.1%
Asian only	2.5%
Native American/Oth Pac I…	0.2%
American Indian/Alaska N…	0.8%
More than one race reported	0.3%

Click the [NA] button in the toolbar to toggle display of missing values:

● **RACE**

Patient race, as recorded by physician. Note it is often left blank.
Use the [NA] toolbar button to toggle how blank values are handled.

Value	Freq
[NA]	--
White only	83.1%
Black/African American only	12.0%
Asian only	3.2%
Native American/Oth Pac I…	0.2%
American Indian/Alaska N…	1.1%
More than one race reported	0.4%

In the latter view, the percentages are normalized over all non-missing values. (Note that this does not affect frequency counts, just the percentages and bar length)

Numeric variables

Protobi automatically bins numeric variables into ranges for analysis. The default is to automatically choose bin sizes based on the standard deviation, and snapping it to a "nice" number that makes sense to those used to counting on two hands of five fingers each:

AGE

Patient age in years.

Value	Freq	
0	2.5%	
1 to 20	16.0%	
21 to 40	18.8%	
41 to 60	29.8%	
61 to 80	26.2%	
81 to 100	6.8%	
Mean	*46.7*	
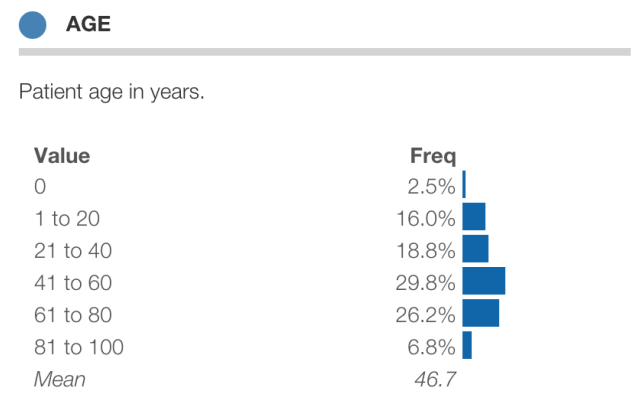
Set bin size

You can change the bin size by clicking on the circle icon ● and selecting "Round by…"

AGE

Copy as text

Copy as image

Round by...

Edit...

	Freq	
	2.5%	
	16.0%	
	18.8%	
41 to 60	29.8%	
61 to 80	26.2%	
81 to 100	6.8%	
Mean	*46.7*	
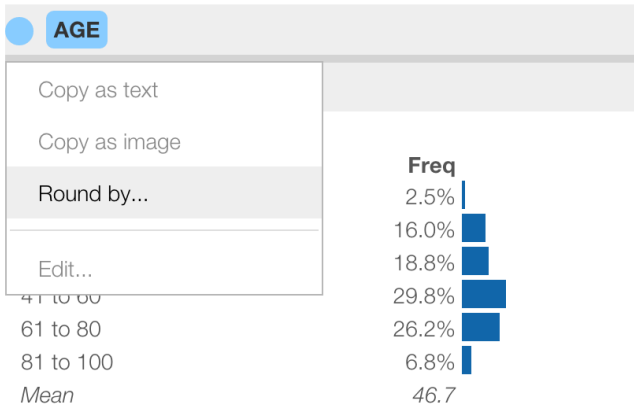
In the dialog you can enter a linear bin size such as '15'. You can also choose 'auto' to automatically choose a bin size, or 'log' to choose logarithmic bins, or choose '0' to show the individual values:

Round by

Enter a bin size for rounding:

number or 'auto'

Ok

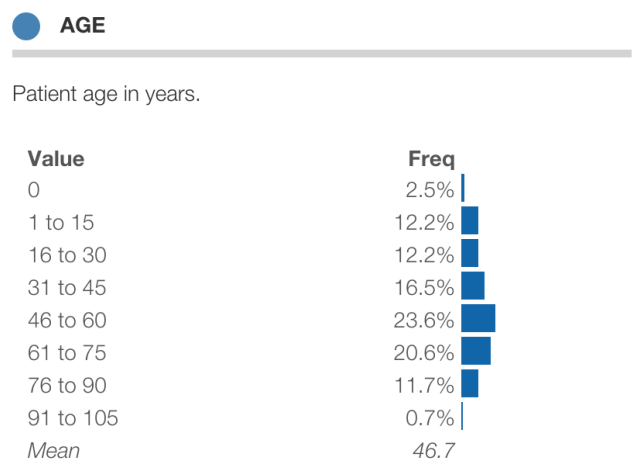Here is the distribution of age with bin size set to 15:

● **AGE**

Patient age in years.

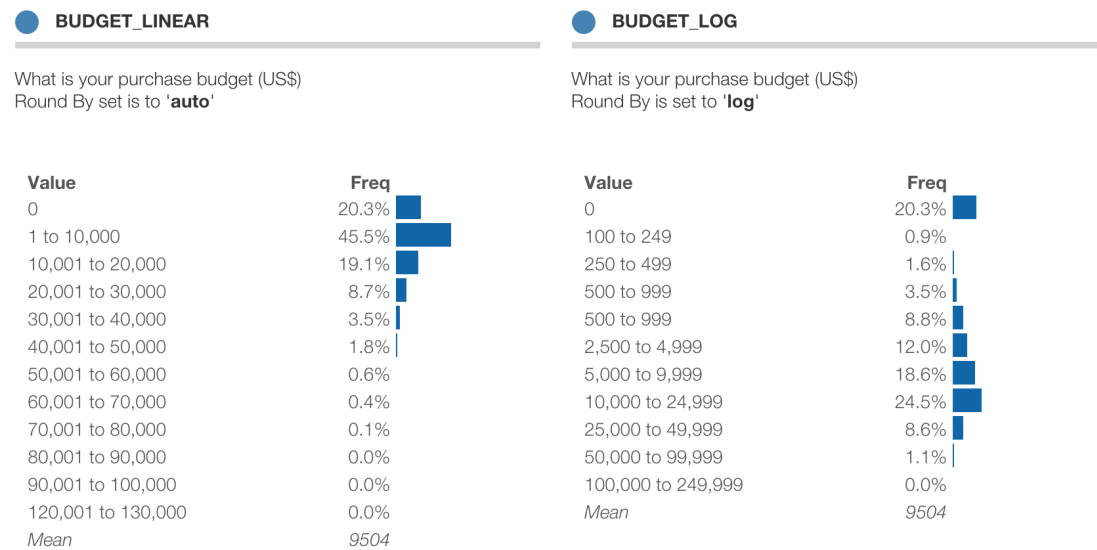| Value	Freq	
0	2.5%	
1 to 15	12.2%	
16 to 30	12.2%	
31 to 45	16.5%	
46 to 60	23.6%	
61 to 75	20.6%	
76 to 90	11.7%	
91 to 105	0.7%	
Mean	*46.7*	

Linear vs Logarithmic Ranges

Many variables such as income, book sales, frequent flier miles, etc. have "heavy-tail" or "long-tail" distributions. These are distributions where the range of values spans multiple orders of magnitude, with many smaller values at the lower end of the range but also many large numbers.

Linear bin ranges make sense for variables that are normally distributed. But for long-tail distributions, the standard deviation is so large as to obscure important differences at the low end. For long-tail distributions, logarithmic ranges can make more sense, compressing the larger values into a bell-shape.

Below is a distribution for income shown with linear ranges (left) and logarithmic ranges (right):

● **BUDGET_LINEAR**

What is your purchase budget (US$)
Round By set is to '**auto**'

Value	Freq	
0	20.3%	
1 to 10,000	45.5%	
10,001 to 20,000	19.1%	
20,001 to 30,000	8.7%	
30,001 to 40,000	3.5%	
40,001 to 50,000	1.8%	
50,001 to 60,000	0.6%	
60,001 to 70,000	0.4%	
70,001 to 80,000	0.1%	
80,001 to 90,000	0.0%	
90,001 to 100,000	0.0%	
120,001 to 130,000	0.0%	
Mean	*9504*	

● **BUDGET_LOG**

What is your purchase budget (US$)
Round By is set to '**log**'

Value	Freq	
0	20.3%	
100 to 249	0.9%	
250 to 499	1.6%	
500 to 999	3.5%	
500 to 999	8.8%	
2,500 to 4,999	12.0%	
5,000 to 9,999	18.6%	
10,000 to 24,999	24.5%	
25,000 to 49,999	8.6%	
50,000 to 99,999	1.1%	
100,000 to 249,999	0.0%	
Mean	*9504*	

Protobi explicitly breaks out '0'and negative values, for which the logarithm is undefined.

Multiple response sets

Often questions come as collections of related items on a common scale. This is common for ratings, rankings "check-all-that-apply" and other kinds of questions.
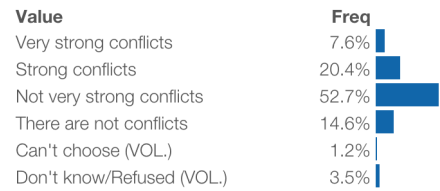
In this survey, **q5** is a set of responses to the question "In your opinion, in AMERICA, how much conflict is there between…". Respondents rated each question on a scale from "Very strong conflicts" to "Don't know". The individual questions Q5a, Q5b, Q5c, Q5d, and Q5e are organized into a group **q5**. By default, each question is displayed as its own independent distribution.

q5

In your opinion, in AMERICA, how much conflict is there between... [Very strong conflicts, strong conflicts, not very strong conflicts, or there are not conflicts?]
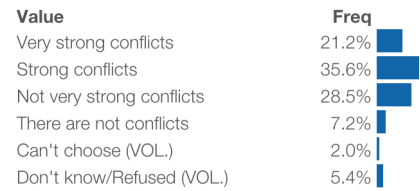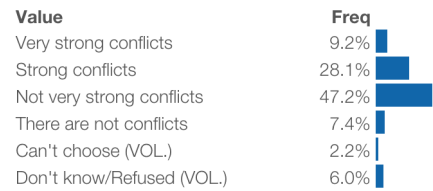
q5a

Q.5a - Young people and older people -

Value	Freq	
Very strong conflicts	7.6%	
Strong conflicts	20.4%	
Not very strong conflicts	52.7%	
There are not conflicts	14.6%	
Can't choose (VOL.)	1.2%	
Don't know/Refused (VOL.)	3.5%	

q5b

Q.5b - Poor people and rich people -

Value	Freq	
Very strong conflicts	21.2%	
Strong conflicts	35.6%	
Not very strong conflicts	28.5%	
There are not conflicts	7.2%	
Can't choose (VOL.)	2.0%	
Don't know/Refused (VOL.)	5.4%	

q5c

Q.5c - Blacks and whites -

Value	Freq	
Very strong conflicts	9.2%	
Strong conflicts	28.1%	
Not very strong conflicts	47.2%	
There are not conflicts	7.4%	
Can't choose (VOL.)	2.2%	
Don't know/Refused (VOL.)	6.0%	

q5d

Q.5d - Immigrants and people born in the United States -

Value	Freq	
Very strong conflicts	15.5%	
Strong conflicts	36.8%	
Not very strong conflicts	32.7%	
There are not conflicts	6.5%	
Can't choose (VOL.)	2.3%	
Don't know/Refused (VOL.)	6.2%	
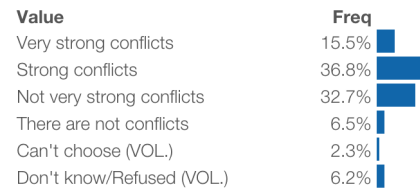
Compact display:

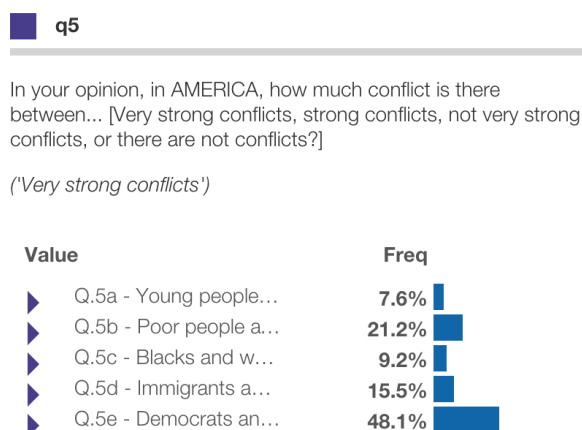These can be displayed in compact form in a number of ways. You can compact the collection variables to show just the mean, the sum or the top-box percentages.

To compact a group to to-box percentages, click the square icon ■ for the group and select "Compact to: 'Very strong conflicts'", as shown below. To expand the display back to showing distributions for each element, select "Expand" from the menu.

q5

Copy as text	...y much conflict is there
Copy as image	...strong conflicts, not very strong
Round by...	...s?]
Expand	
Compact to mean	
Compact to 'Very strong confl...	...people -
Compact to 'Don't know/Refu...	
Compact to ...	
Condense	
Edit...	
Extract common title text...	
Extract titles from formats...	

q5b

Q.5b - Poor people and rich people -

	Freq
7.6%	
20.4%	
52.7%	
14.6%	
1.2%	
3.5%	

Value	**Freq**
Very strong conflicts	21.2%
Strong conflicts	35.6%
Not very strong conflicts	28.5%
There are not conflicts	7.2%
Can't choose (VOL.)	2.0%
Don't know/Refused (VOL.)	5.4%

q5c

Q.5c - Blacks and whites -

Value	**Freq**
Very strong conflicts	9.2%
Strong conflicts	28.1%
Not very strong conflicts	47.2%
There are not conflicts	7.4%
Can't choose (VOL.)	2.2%
Don't know/Refused (VOL.)	6.0%

q5d

Q.5d - Immigrants and people born in the United States -

Value	**Freq**
Very strong conflicts	15.5%
Strong conflicts	36.8%
Not very strong conflicts	32.7%
There are not conflicts	6.5%
Can't choose (VOL.)	2.3%
Don't know/Refused (VOL.)	6.2%

The distributions are now concisely shown as just the top-box percentages for the specified value.

q5

In your opinion, in AMERICA, how much conflict is there between... [Very strong conflicts, strong conflicts, not very strong conflicts, or there are not conflicts?]

('Very strong conflicts')

Value	**Freq**
▶ Q.5a - Young people…	**7.6%**
▶ Q.5b - Poor people a…	**21.2%**
▶ Q.5c - Blacks and w…	**9.2%**
▶ Q.5d - Immigrants a…	**15.5%**
▶ Q.5e - Democrats an…	**48.1%**

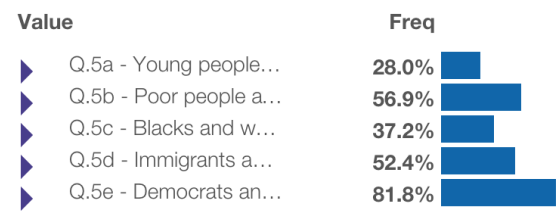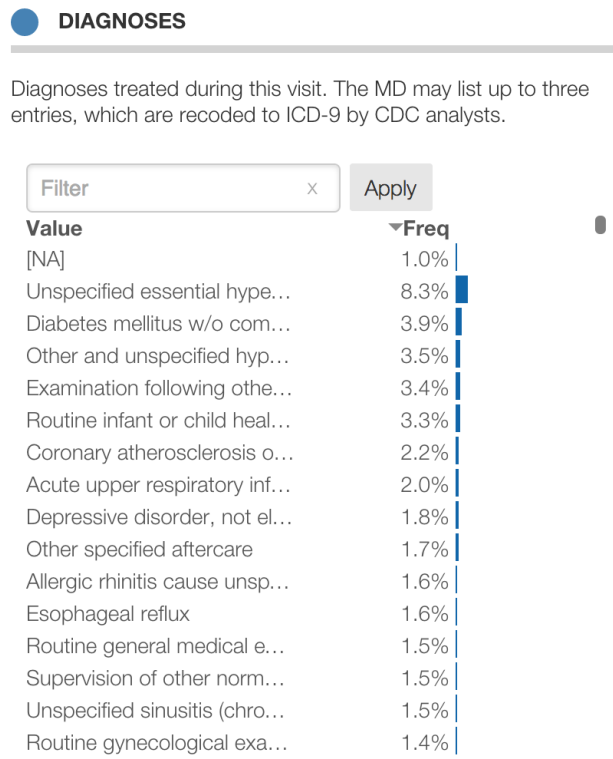You can click these as any binary variable. Click the triangle icon to see the entire distribution.

Protobi shows the first and last category as options. You can select others by choosing "Compact to…" and entering specific values to compact to. Recognized values include:

- Unformatted value, e.g. `1`
- Multiple unformatted values, e.g. `[1,2]`
- Mean, `$mean`
- Sum, `$sum`

Below shows the top-two-box percentages by specifying `[1,2]` under "Compact to…":

q5

In your opinion, in AMERICA, how much conflict is there between... [Very strong conflicts, strong conflicts, not very strong conflicts, or there are not conflicts?]

('Very strong conflicts', or 'Strong conflicts')

Value	Freq
▶ Q.5a - Young people…	**28.0%**
▶ Q.5b - Poor people a…	**56.9%**
▶ Q.5c - Blacks and w…	**37.2%**
▶ Q.5d - Immigrants a…	**52.4%**
▶ Q.5e - Democrats an…	**81.8%**

Text verbatims

Protobi can handle text verbatim responses elegantly. Here MDs were asked to specify up to three diagnoses for each patient visit in a given period. Collectively this spans thousands of diagnoses for tens of thousands of patient visits.

DIAGNOSES

Diagnoses treated during this visit. The MD may list up to three entries, which are recoded to ICD-9 by CDC analysts.

Filter	x	Apply

Value	▼Freq	
[NA]	1.0%	
Unspecified essential hype…	8.3%	
Diabetes mellitus w/o com…	3.9%	
Other and unspecified hyp…	3.5%	
Examination following othe…	3.4%	
Routine infant or child heal…	3.3%	
Coronary atherosclerosis o…	2.2%	
Acute upper respiratory inf…	2.0%	
Depressive disorder, not el…	1.8%	
Other specified aftercare	1.7%	
Allergic rhinitis cause unsp…	1.6%	
Esophageal reflux	1.6%	
Routine general medical e…	1.5%	
Supervision of other norm…	1.5%	
Unspecified sinusitis (chro…	1.5%	
Routine gynecological exa…	1.4%	

By default, Protobi sorts text responses in by frequency (rather than value) when there are more than 12 values and the values are either formatted or not numeric.

Search verbatims

You can screen values for visibility by typing in the "filter" input. Here only values that have the text "lupus" are shown, which is either "**Lupus** erythematosus" or "Systemic **lupus** erythematosus":

DIAGNOSES

Diagnoses treated during this visit. The MD may list up to three entries, which are recoded to ICD-9 by CDC analysts.

lupus ✕	Apply

Value	▼Freq
Systemic lupus erythemat…	0.1%
Lupus erythematosus	0.0%

Searches as filters

Click the "Apply" button to make this search an actual filter which subsets the data and updates distributions for all the other elements. Note that MDs were allowed to enter multiple diagnoses. The distribution below includes only patients who have a diagnosis including "Lupus" but they may also have other diagnoses too:

DIAGNOSES

Diagnoses treated during this visit. The MD may list up to three entries, which are recoded to ICD-9 by CDC analysts.

Filter ✕	Apply

Value	▼Freq
[NA]	0.0%
Systemic lupus erythemat…	79.2%
Lupus erythematosus	20.8%
Rheumatoid arthritis	8.3%
Mylagia and myositis, uns…	8.3%
Unspecified hypothyroidism	8.3%
Arteritis, unspecified	8.3%
Long-term (current) use of …	8.3%
Adhesive capsulitis of sho…	4.2%
Pain in joint of forearm	4.2%
Pain in joint of shoulder re…	4.2%
Other and unspecified non…	4.2%
Acute sinusitis, unspecified	4.2%
Arthropathy, unspecified, s…	4.2%
Nephritis/nephropathy, not…	4.2%
Unspecified essential hype…	4.2%

Regex search

Protobi search is very powerful as it is based on Regular Expressions. For instance:

- `lupus` matches any element containing the string `lupus`
- `itis$` matches any element ending with the string `itis$`
- `^lu` matches any string starting with `lu`
- `itis|algia` matches any string containing `itis` or `algia

Below is a search for `^lu` which matches entries that begin with `lu`, and thus matches "**Lu**pus erythematosus" but does not match "Systemic lupus erythematosus":

DIAGNOSES

Diagnoses treated during this visit. The MD may list up to three entries, which are recoded to ICD-9 by CDC analysts.

^lu	✕	Apply

Value	▼Freq
Lumbago	1.0%
Lumbar intervertebral disc …	0.2%
Lump or mass in breast	0.1%
Lumbosacral spondylosis …	0.1%
Lupus erythematosus	0.0%
Lumbar spinal cord injury	0.0%
Lumbosacral plexus lesions	0.0%
Luxation of globe	0.0%
Lumbosacral root lesions, …	0.0%
Lung involvement in other …	0.0%

Multiple verbatims: Expanded vs Condensed

Sometimes surveys will ask respondents for multiple open ended answers. For instance, "What are the key strengths of this new product? Please list up to three".

Rather than see them as three separate fields, we might wish to combine them into one multiple-response category. You can do that by putting the elements into a single group. Then click the square icon for the group and select "Condense" from the context menu. This "squishes" the data horizontally into a single multi-response variables.

Here's an example using Diagnoses from the CDC NAMCS survey, where MDs list up to three diagnoses for each patient. In this view, we've drill into patients with diabetes to highlight that a record may have more than one value at a time, so the percentages sum to more than 100%:

DIAGNOSES

Diagnoses treated during this visit. The MD may list up to three entries, which are recoded to ICD-9 by CDC analysts.

| Filter | x | Apply |

Value	▼Freq	
[NA]	0.0%	◁
Diabetes mellitus w/o com…	100.0%	
Unspecified essential hype…	39.6%	▶
Other and unspecified hyp…	16.9%	▶
Examination following othe…	4.2%	
Coronary atherosclerosis o…	4.0%	▶
Benign essential hypertens…	3.1%	▶
Pure hypercholesterolemia	3.0%	▶
Unspecified hypothyroidism	2.6%	▶
Obesity, unspecified	2.5%	▶
Atrial fibrillation	2.4%	▶
Chronic airway obstruction…	1.9%	▶
Esophageal reflux	1.5%	
Congestive heart failure	1.4%	▶
Depressive disorder, not el…	1.3%	
Unspecified cataract	1.3%	

- **Clicking** a condensed element selects all records where **any** of the elements match.
- **Shift+Clicking** a condensed element selects all records where **none** of the elements match.
- A condensed element is blank ([na]) only if **all** the values are blank

Crosstabs

Protobi allows you to easily crosstab elements against each other.

Individual crosstabs

To create an individual crosstab, drag one variable which defines the columns (or "banner") and drop it onto another variable which will will define the rows (or "stub"). In the example below, **q2** is dropped onto **q1** to create this crosstab:

● **q1 vs. q2**

Q.1 Generally, how would you say things are these days in your life - would you say that you are very happy, pretty happy, or not too happy?

q1	Overall	Excellent	Good	Only fair	Poor	Don't know/ Refused (VOL.)
		q2: Q.2 How would you rate your own health in general these days? Would you say your health is excellent, good, only fair, or poor?				
Very happy	30.5%	▲49.6%	▼27.6%	▼18.2%	▼7.4%	28.6%
Pretty happy	50.9%	▼42.7%	▲57.8%	48.4%	▼31.1%	--
Not too happy	15.4%	▼6.0%	▼11.3%	▲29.6%	▲54.9%	42.9%
Don't know/Refused (VOL.)	3.2%	▼1.7%	3.3%	3.8%	▲6.6%	28.6%
N	2,511	637	1,299	446	122	7

Here we can see that of respondents who entered 'Extremely' to **q2**, 49.6% answered "Very happy" to **q1**.

Shading and arrows indicate statistically significant differences from the set of all *other* columns.

You can create crosstabs using collections of variables as the stub. For instance below we can see that respondents who answer "Not to happy" in **q1** are significantly more likely to select "Very strong conflicts" for many of the subquestions in **q5**:

In your opinion, in AMERICA, how much conflict is there between...

('Very strong conflicts')

q5	Overall	Very happy	Pretty happy	Not too happy	Don't know/ Refused (VOL.)
		q1: Q.1 Generally, how would you say things are these days in your life - would you say that you are very happy, pretty happy, or not too happy?			
N	2,511	766	1,277	387	81
Q.5a - Young people and …	7.6%	7.0%	▼6.0%	▲13.2%	12.3%
Q.5b - Poor people and ric…	21.2%	19.3%	20.4%	▲28.7%	17.3%
Q.5c - Blacks and whites -…	9.2%	8.4%	8.2%	▲14.0%	8.6%
Q.5d - Immigrants and pe…	15.5%	14.9%	▼13.7%	▲23.3%	13.6%
Q.5e - Democrats and Re…	48.1%	49.6%	47.1%	50.4%	38.3%

Any element can be used to define the rows (or "stubs") in a crosstab. Only individual fields (i.e. the ones with circle icons) can be used to define crosstab columns (or "banner")

Global crosstabs

You can crosstab all the elements in the research

Export

There are several ways to export data from Protobi. Protobi is principally designed as a tool to find insights and interesting subsets, and makes it easy to get data into PowerPoint for a quick summary or into Excel for more detailed charting and analysis.

As image (e.g. to PowerPoint)

Click the circle icon and select "Copy as image". This will bring up a dialog with the element as an image. Right click and select "Copy as image…" to copy it to your clipboard or "Save image as…" to save it to your computer.
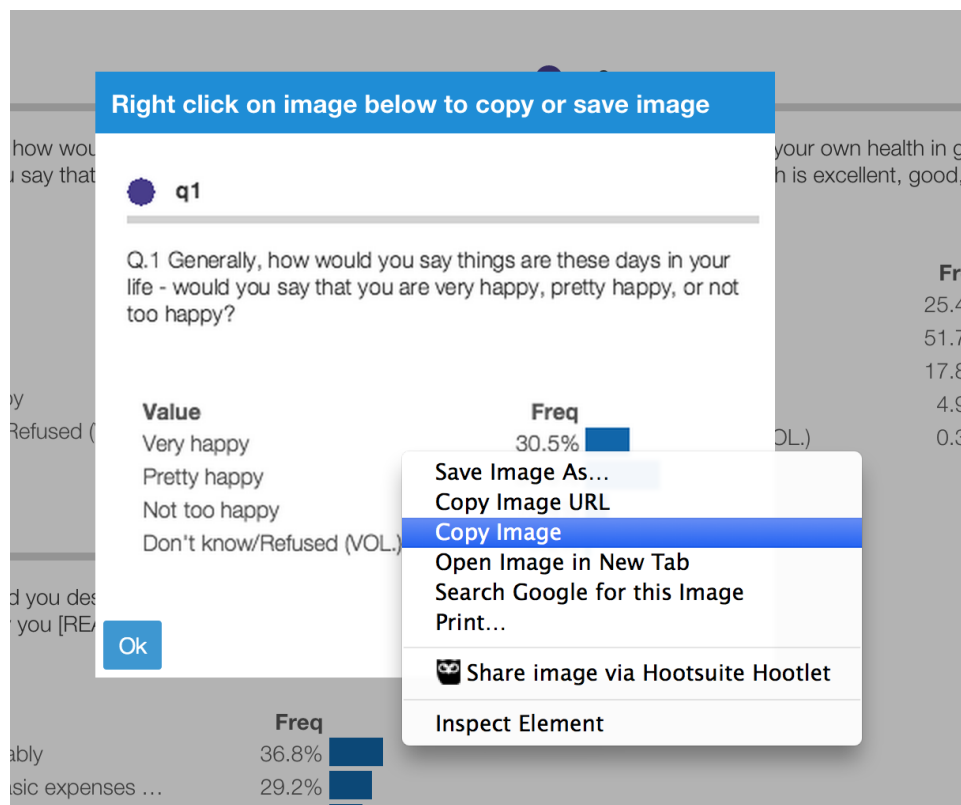
This two-step process is necessary as browser security restrictions prevent programs from putting information onto your clipboard or computer without your direct action.

As PDF (e.g. to print)

You can save the entire project as a PDF by selecting "Print…" from the browser's File menu, and selecting "Print to PDF". **Note:** This works best in Chrome and Firefox.

As text (e.g. to Excel)

Alternative, click the circle icon and select "Copy as text". This will select the element in the browser:
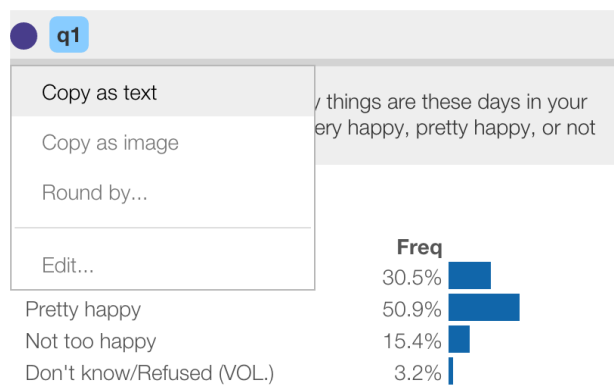
You can then use the browser to copy/paste the data to Excel (or use the Ctrl+C / Ctrl+V keyboard shortcuts) to create your own custom charts:

Create a new project

You can create a new Protobi project simply by importing a datafile.

Click the "New project" widget under your welcome screen and use the File dialog to select a SAV or CSV datafile. Or drag-and-drop the data file from Windows Explorer or Mac OSX Finder on to the "New project" widget.

For this example, we'll use the Gender and Generations survey which you can download from Pew Research at http://www.pewsocialtrends.org/2014/01/07/gender-and-generations/ . The download includes a formatted SPSS datafile

Nov12 Gender and
Generations.sav

After a few seconds, the new project will appear as a new widget in your welcome page, and automatically open.

Nov12 Gender and Generations

Description
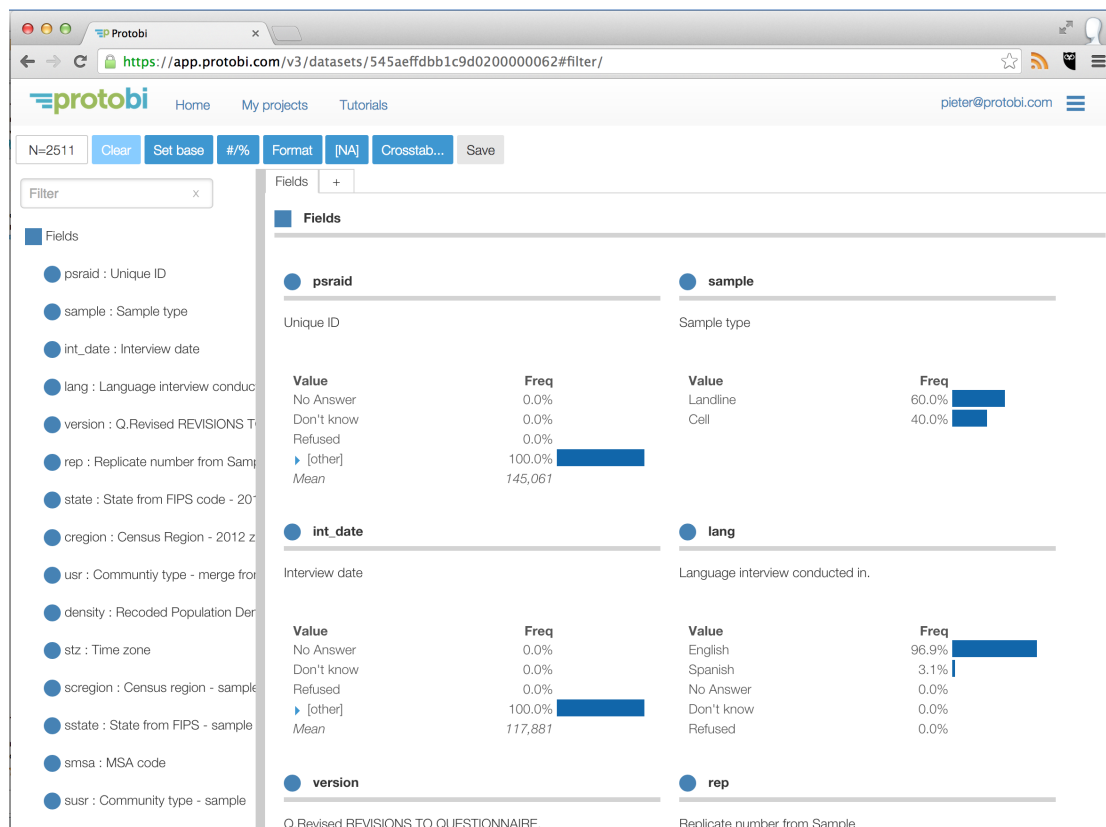Nov12 Gender and Generations.sav

Created
Today at 10:50 PM

Data file
Nov12 Gender and Generations_…

▶ Open   👤 Share   ✖ Delete

When you first create a project, all the fields are placed in a group called "Fields" which is also the main tab:

Voila! You now have an interactive view of all the fields in your dataset. It has the variable labels and value formats and other hints from the SPSS file. Numeric fields are automatically binned. You can start exploring!

Recognized datafiles

Protobi reads rectangular "flat" datafiles, that have rows representing observations and columns representing attributes. It can currently read two formats:

- SPSS datafiles with ".sav" extensions,
- Comma-Separated Value files with ".csv" extensions

CSV files must have a first row with column names. It must also be RFC 4180 IETF compliant, which means that text values with embedded commas, line breaks or quotes must be enclosed in double quotes, and any double quotes must be escaped as double double-double quotes. Fortunately, this is the standard CSV that Excel and most applications use.

SPSS datafiles are great for many reasons. They are standard output from most survey platforms. The datafile includes a lot of metadata, including

- the variable type (i.e. String or Number),
- variable labels
- value formats
- weight field
- multi-reponse sets

If you have the option of receiving your data in SPSS format, that will make it very easy to work with Protobi.

Organize

When you first create a project, all the fields are placed in a group called "Fields". You can use this view immediately, and that might be enough. Often you'll wish to organize it a bit further.

Create a new tab

Click the + icon on the row of tabs to create new tab. You'll be prompted to provide a name, let;s call it "Intro". You can drag elements into this view from the left:
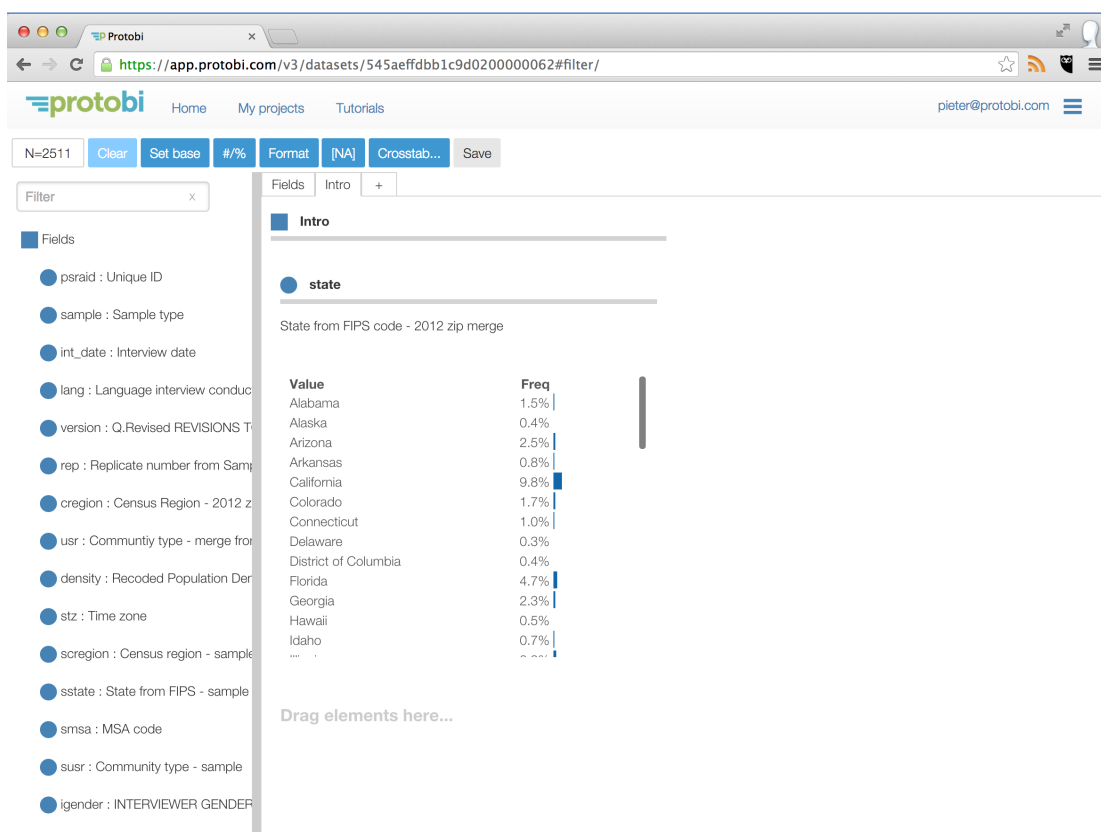
Group/ungroup elements

Survey datasets often have structure to them, and you may wish to organize the elements into groups to reflect that.

In this example, scroll down in the Tree to questions 'q5a', 'q5b', 'q5c', 'q5d', and 'q5e'. We might wish to place these into a group, say **q5**.

To create a group, select them by clicking on them in the Tree. Then use the keyboard shortcut "Shift+G".

Protobi prompts you for the name of the group and suggests one using the longest common prefix among the selected elements. Here it suggests 'q5' which we'll accept.

Autogroup

Surveys often have not just one group to define, but dozens or even hundreds. Protobi can intelligently autogroup elements based on their name.

Select the group from the Panel, in this case "Fields", click the square icon and select "Autogroup" from the context menu:
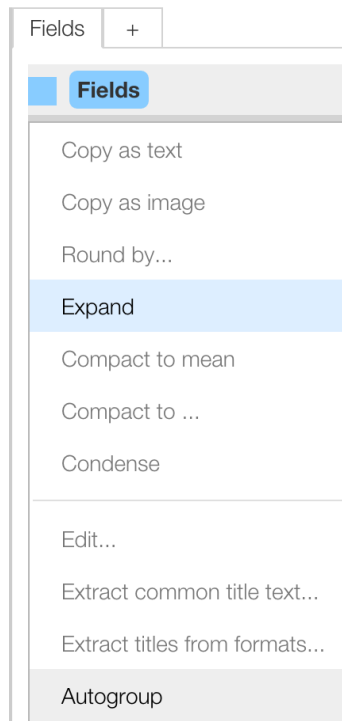
This may take a while, but after a few seconds or maybe 60 seconds will yield a message box indicating success. The elements are now organized into groups based on careful parsing of the name, intelligently sorting numerals numerically and other characters alphabetically.

The results for this example are below (these appear after scrolling a bit down to the elements starting with 'q'), as these are now arranged alphabetically):
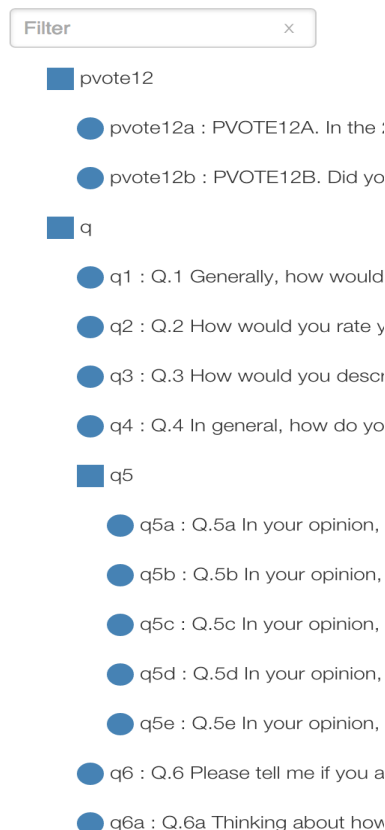
Tip: *Now is a good time to click the "Save" button to keep your work!*

Drag/drop elements

You can drag-drop elements:

- within the Tree to reorder and regroup them
- to the Panel to place them in a tab (which also regroups them)
- to the tab list at the top of the Panel to make them a tab (and a top-level group)

Here we dragged the element **q** to the tab panel area (after the '+', highlighted in gold) to make it a new tab. Note that this also makes it a top-level group in the Tree, parallel with 'Fields':
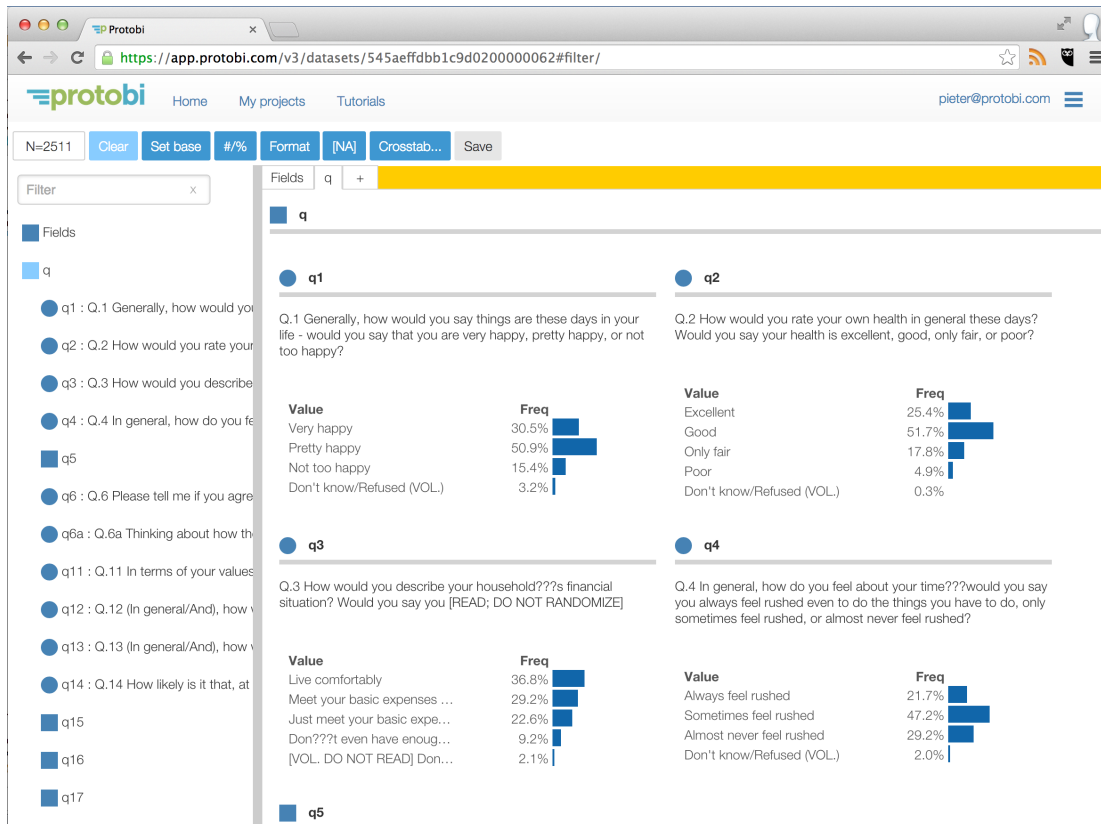
Edit titles

As an editor, you can edit the title text of elements by clicking on it. It will automatically turn into an editable text area.

For instance, the survey had a character in **q3** that got converted to '???' somewhere in creating the data file. You can change that to an apostrophe by clicking and editing.

● **q3**

Q.3 How would you describe your household???s financial
situation? Would you say you [READ; DO NOT RANDOMIZE]

Value	Freq
Live comfortably	36.8%
Meet your basic expenses …	29.2%
Just meet your basic expe…	22.6%
Don???t even have enoug…	9.2%
[VOL. DO NOT READ] Don…	2.1%

The change is immediate, but only within your session. To make it permanent and visible to other users, hit the "Save" button.

Note: To change the key that identifies the element (e.g. "q3"), use the Advanced Properties editor (described below) to change the **displayKey** property.

Extract common title text from groups

A common pattern in surveys is for collections of related questions to appear in a battery with common text. While the individual subquestions have data values and get columns in the data file, the common text isn't associated with a data field. So the survey platform will include that text in the variable labels for the subquestions.

For instance, in q5 we see that they all have the common text "In your opinion, in AMERICA, how much conflict is there between… " and "…Very strong conflicts, strong conflicts, not very strong conflicts, or there are not conflicts?"
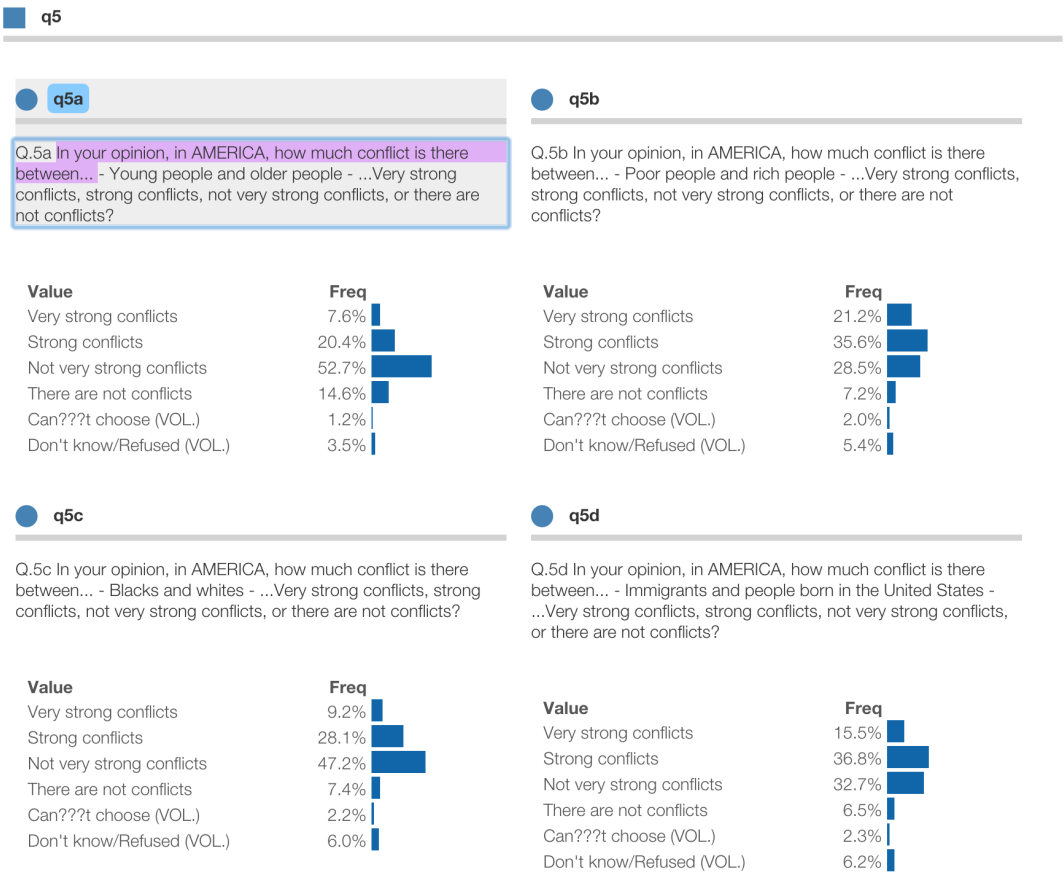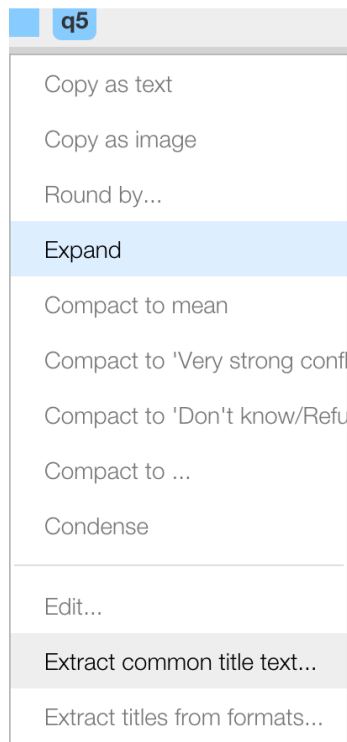
q5

q5a

Q.5a In your opinion, in AMERICA, how much conflict is there between... - Young people and older people - ...Very strong conflicts, strong conflicts, not very strong conflicts, or there are not conflicts?

Value	Freq
Very strong conflicts	7.6%
Strong conflicts	20.4%
Not very strong conflicts	52.7%
There are not conflicts	14.6%
Can???t choose (VOL.)	1.2%
Don't know/Refused (VOL.)	3.5%

q5b

Q.5b In your opinion, in AMERICA, how much conflict is there between... - Poor people and rich people - ...Very strong conflicts, strong conflicts, not very strong conflicts, or there are not conflicts?

Value	Freq
Very strong conflicts	21.2%
Strong conflicts	35.6%
Not very strong conflicts	28.5%
There are not conflicts	7.2%
Can???t choose (VOL.)	2.0%
Don't know/Refused (VOL.)	5.4%

q5c

Q.5c In your opinion, in AMERICA, how much conflict is there between... - Blacks and whites - ...Very strong conflicts, strong conflicts, not very strong conflicts, or there are not conflicts?

Value	Freq
Very strong conflicts	9.2%
Strong conflicts	28.1%
Not very strong conflicts	47.2%
There are not conflicts	7.4%
Can???t choose (VOL.)	2.2%
Don't know/Refused (VOL.)	6.0%

q5d

Q.5d In your opinion, in AMERICA, how much conflict is there between... - Immigrants and people born in the United States - ...Very strong conflicts, strong conflicts, not very strong conflicts, or there are not conflicts?

Value	Freq
Very strong conflicts	15.5%
Strong conflicts	36.8%
Not very strong conflicts	32.7%
There are not conflicts	6.5%
Can???t choose (VOL.)	2.3%
Don't know/Refused (VOL.)	6.2%

Protobi can extract that common text into the parent group. Click the square icon for the group and select "Extract common text…" from the context menu.

q5

Copy as text

Copy as image

Round by...

Expand

Compact to mean

Compact to 'Very strong confli

Compact to 'Don't know/Refu

Compact to ...

Condense

Edit...

Extract common title text...

Extract titles from formats...

Protobi will prompt you for the text to extract, and suggest a guess based on the longest common prefix. Here they don't have a common prefix, so highlight the text yourself, and copy/paste it to the dialog:

The page at https://app.protobi.com says:

Extract the following text from each element:

In your opinion, in AMERICA, how much conflict is

Cancel   OK
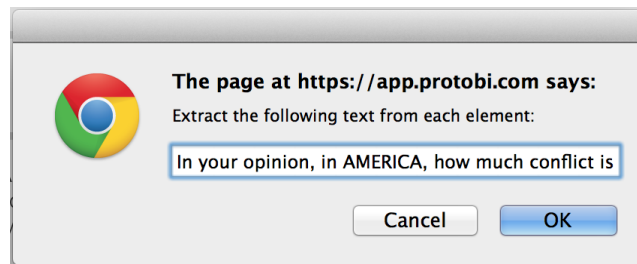
Protobi removes that text from each element and places it in the parent group:

q5

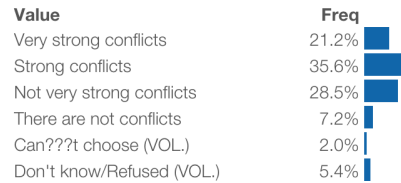In your opinion, in AMERICA, how much conflict is there between...

q5a

Q.5a - Young people and older people - ...Very strong conflicts, strong conflicts, not very strong conflicts, or there are not conflicts?

Value	Freq
Very strong conflicts	7.6%
Strong conflicts	20.4%
Not very strong conflicts	52.7%
There are not conflicts	14.6%
Can???t choose (VOL.)	1.2%
Don't know/Refused (VOL.)	3.5%

q5b

Q.5b - Poor people and rich people - ...Very strong conflicts, strong conflicts, not very strong conflicts, or there are not conflicts?

| Value	Freq
Very strong conflicts	21.2%
Strong conflicts	35.6%
Not very strong conflicts	28.5%
There are not conflicts	7.2%
Can???t choose (VOL.)	2.0%
Don't know/Refused (VOL.)	5.4%

q5c

Q.5c - Blacks and whites - ...Very strong conflicts, strong conflicts, not very strong conflicts, or there are not conflicts?

Value	Freq
Very strong conflicts	9.2%
Strong conflicts	28.1%
Not very strong conflicts	47.2%
There are not conflicts	7.4%
Can???t choose (VOL.)	2.2%

q5d

Q.5d - Immigrants and people born in the United States - ...Very strong conflicts, strong conflicts, not very strong conflicts, or there are not conflicts?

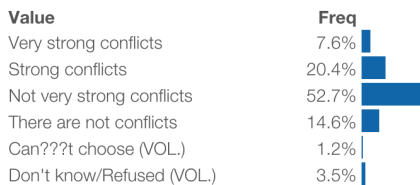| Value	Freq
Very strong conflicts	15.5%
Strong conflicts	36.8%
Not very strong conflicts	32.7%
There are not conflicts	6.5%
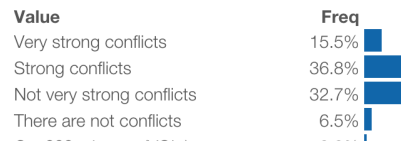
This particular question is a hassle as it has two fragments of common text, separated by variable text. We can repeat the process with the other fragment:

q5

In your opinion, in AMERICA, how much conflict is there between... ...Very strong conflicts, strong conflicts, not very strong conflicts, or there are not conflicts?
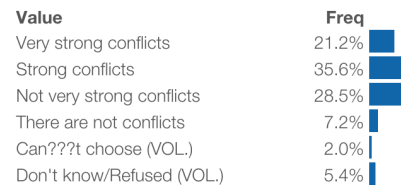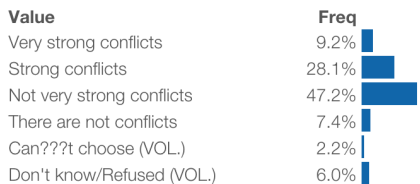
q5a

Q.5a - Young people and older people -

Value	Freq
Very strong conflicts	7.6%
Strong conflicts	20.4%
Not very strong conflicts	52.7%
There are not conflicts	14.6%
Can???t choose (VOL.)	1.2%
Don't know/Refused (VOL.)	3.5%

q5b

Q.5b - Poor people and rich people -

Value	Freq
Very strong conflicts	21.2%
Strong conflicts	35.6%
Not very strong conflicts	28.5%
There are not conflicts	7.2%
Can???t choose (VOL.)	2.0%
Don't know/Refused (VOL.)	5.4%

q5c

Q.5c - Blacks and whites -

Value	Freq
Very strong conflicts	9.2%
Strong conflicts	28.1%
Not very strong conflicts	47.2%
There are not conflicts	7.4%
Can???t choose (VOL.)	2.2%
Don't know/Refused (VOL.)	6.0%

q5d

Q.5d - Immigrants and people born in the United States -

Value	Freq
Very strong conflicts	15.5%
Strong conflicts	36.8%
Not very strong conflicts	32.7%
There are not conflicts	6.5%
Can???t choose (VOL.)	2.3%
Don't know/Refused (VOL.)	6.2%
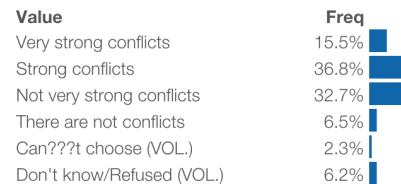
At this point, it might be nice to compact the display to top-box percentages:

q5

In your opinion, in AMERICA, how much conflict is there between... ...Very strong conflicts, strong conflicts, not very strong conflicts, or there are not conflicts?

('Very strong conflicts')

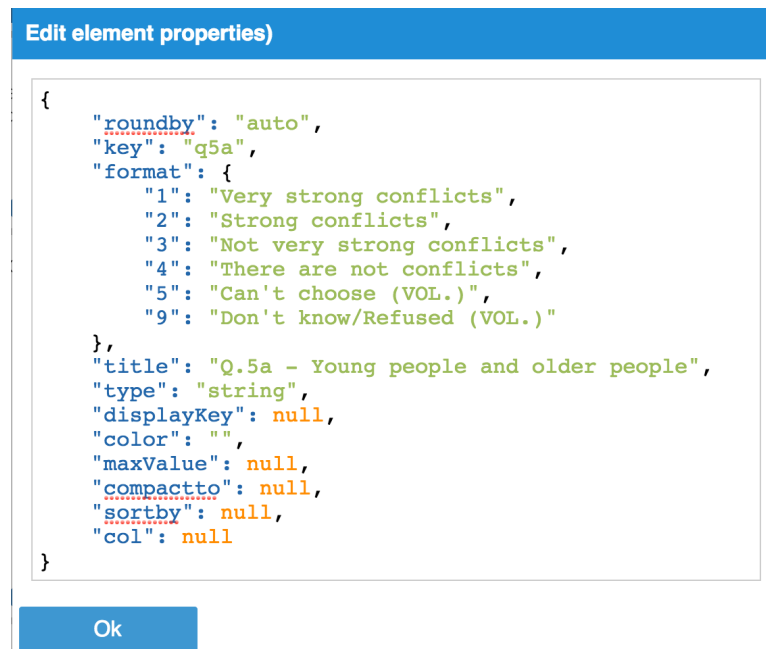| Value	Freq	
Q.5a - Young people…	**7.6%**	
Q.5b - Poor people a…	**21.2%**	
Q.5c - Blacks and w…	**9.2%**	
Q.5d - Immigrants a…	**15.5%**	
Q.5e - Democrats an…	**48.1%**	
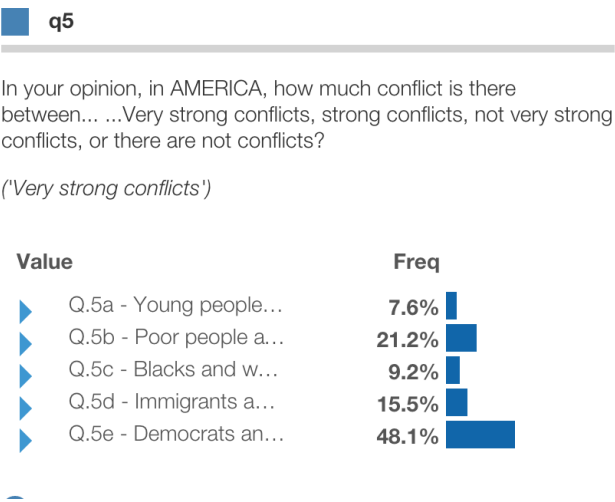
Tip: *Now is a good time to click the "Save" button to keep your work!*

Edit advanced properties

You can edit advanced properties for any element or group by selecting "Edit…" from the context menu. This will bring up an editor, which allows you to modify all the properties directly.

This uses JSON syntax, which has the advantage of handling nearly any property, but requires care. The editor highlights your syntax after you type, to help do it correctly.

```
Edit element properties)

{
    "roundby": "auto",
    "key": "q5a",
    "format": {
        "1": "Very strong conflicts",
        "2": "Strong conflicts",
        "3": "Not very strong conflicts",
        "4": "There are not conflicts",
        "5": "Can't choose (VOL.)",
        "9": "Don't know/Refused (VOL.)"
    },
    "title": "Q.5a - Young people and older people",
    "type": "string",
    "displayKey": null,
    "color": "",
    "maxValue": null,
    "compactto": null,
    "sortby": null,
    "col": null
}
```

Ok

The following properties are recognized:

- **key** unique identifier for element
- **field** name of data field, if different from **key**
- **displayKey** identifier to display, if different from **key**
- **title** longer text description
- **type** type of data, may be "number", "string", or "empty" or null
- **roundby** bin size for numbers, maybe 0, "auto", "log" or a positive number
- **sortby** either "value" or "freq" or null
- **color** any CSS color, either named (e.g. "gold") or hexadecimal (e.g. "#FA0")
- **compactto** a single raw value (e.g. 5), an array (e.g. [4,5]), "$mean" or "$sum"
- **maxValue** to scale bar lengths when compacting to $mean, e.g. 100 for percentages
- **format** map of raw values to value formats, e.g. {"1": "Strongly agree", …}`
- **children** array of keys defining a group, e.g. ["q5a", "q5b", "q5c", "q5d"]

Share URL

You can share the URL to a project. The URL encodes the current scenario as you click, so pasting the URL into the browser recovers your application state.

For instance, the dataset used for many of the examples in this tutorial can be found at: http://app.protobi.com/v3/datasets/53878aa0618c90020000015c

If your project is public, then anyone with the URL can access the project. If the project is private, the user must be signed in and specified as a permitted user.

You have the ability to open, share and delete projects that you create.

Add colleagues and manage permissions

If you have Admin privileges to the project (which is the default for projects you create), you can mange user permissions. Click the trigram icon in the upper right of the project application and select "Project permissions…"
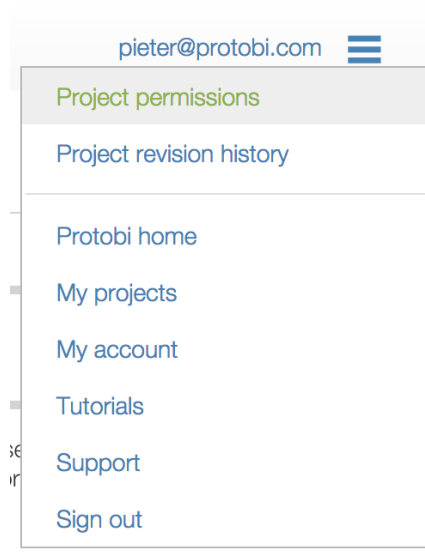
There are four levels of permission:

- **View** users can open, query and explore the data.
- **Edit** users can also edit the map and save changes for all users.
- **Admin** users can manage user permissions and project visibility.
- **None** removes user from permissions for this project.

You will need to notify your colleagues that you have added them to the project. Protobi does not automatically email users when permissions change.

Try it

Try Protobi yourself, using a dataset of your own, or the car_sales.sav example dataset:box-sizing: border-box;

1. **Download** the *car_sales.sav* datafile from http://protobi.com/data/sav/car_sales.sav

car_sales.sav

1. **Sign in** to Protobi at https://app.protobi.com

2. **Create a new project** by dragging the above file to the "New project" widget

3. **Group** all the standardized "z" variables into a group called "Z"

4. **Create a tab** by dragging the group "Z" from the Tree to the tabs above the Panel

5. **Find an element** "Type" in the Panel by looking in the tab "Fields"

6. **Toggle display** from percentages to counts by pressing the "#/%" button in the Toolbar

7. **Toggle display** of formatted values by pressing the "Format" button in the Toolbar

8. **Create a filter** by clicking "Trucks" within the element "Type"

9. **Find significant differences** between cars and trucks by scanning

10. **Crosstab** vehicle MPG by vehicle Type by dragging "Type" onto "Vehicle"

11. **Crosstab** all elements by dragging "Type" to the Crosstab button in the Toolbar

12. **Change bins** for "MPG" by clicking the circle icon and selecting "Round by…"

13. **Edit format** for the element "Type", perhaps change "Truck" to "Truck/SUV"

14. **Compact** the group "Z" to the mean (these are standardized, so the means are zero)

15. **Copy** and element as an Image to PowerPoint

16. **Copy** an element as Text to Excel

Support and Training

We like to help you make the most of Protobi for your data, and make it part of your favorite toolset. If you have questions or ideas, please contact us at support@protobi.com.