

Release notes for July 2017 StatCrunch updates

Enhancements:

- A new **Featured data sets** page has been added with an initial inventory of 10 featured data sets. See page 2 for details.
- The result history has been improved with more details provided for result titles. See page 3 for details.
- A new **Randomization test for slope** applet was added that allows for resampling while tracking the slope between two variables. See page 4 for details.

Minor fixes/enhancements:

- All applets that simulate data were given an *Analyze* button that will export the simulated results into the data table.
- A second version of the Geometric distributional calculator is now available that will calculate the number of trials before the first success.
- The account renewal process has been improved.
- A new expression has been added to calculate the proportion of a column that equals a specified value: `prop(colName, value)`.
- Various usability improvements were implemented to the statcrunch.com pages along with the StatCrunch application.
- Miscellaneous bugs with StatCrunch and its webpages were corrected.





Featured Data Sets Page

StatCrunch has always provided a robust list of data sets shared through the community of StatCrunch users. Currently there are over 30,000 shared data sets available in StatCrunch at the [Shared data page](#). In addition to the shared data sets is a new [Featured data page](#) which contains an abbreviated list of high quality, real-world data sets.

The initial featured data sets list contains 10 real data sets that cover some high interest topics. Over time this list will continue to grow as new data sets will be added. These data sets include a detailed description and have been tagged for easy searching based on statistical topics.

Below is a screenshot of the new [Featured data page](#). As new data sets are added, they will appear at the top of this page and will likewise appear on the homepage.

The screenshot shows the StatCrunch website interface. The top navigation bar includes 'Home', 'Explore', 'MyStatCrunch', 'Open StatCrunch', 'Resources', and 'Support'. The 'Explore' menu is open, and 'Featured Data' is highlighted. Below the navigation bar, there is a search bar and a list of popular data tags. The main content area is titled 'Featured data sets' and shows a table of 12 data sets. The first four data sets are visible in the screenshot:

Data Set/Description	Last edited	Size	Views										
 Shark Attacks Worldwide This data comes from www.sharkattackfile.net. It records data on all shark attacks in recorded history including attacks before 1800. Included is all known information on the shark attack including the date, location, information on the individual who was attacked, details on the injuries sustained by the victim, and the species of the shark	Apr 3, 2017	1MB	2										
 Movie Budgets and Box Office Earnings (Updated Fall 2016) This data all comes from the following website that tracks the financial performance of movies: http://www.the-numbers.com/movie/budgets/all The "Budget", "Domestic Gross", and "Worldwide Gross" columns each are in millions of dollars.	Apr 3, 2017	266KB	3										
 California Home Prices, 2009 This dataset is a collection of real estate listings from San Luis Obispo county, California, and some locations around it from 2009. The prices are their list price at the creation of this dataset. For more information about this data, go to the website source listed above.	Apr 3, 2017	46KB	4										
 US Workforce Participation This data primarily comes from two sources: Federal Reserve Bank of St. Louis and the US Bureau of Labor Statistics . <table border="1"> <thead> <tr> <th>Column</th> <th>Description</th> </tr> </thead> <tbody> <tr> <td>Year</td> <td>The calendar year for each value</td> </tr> <tr> <td>Annual Average Workforce Participation</td> <td>Defined by the Bureau of Labor Statistics as "the percentage of the population [16 years and older] that is either employed or unemployed (that is, either working or actively seeking work). Note that 2015's Annual Average is calculated using the first 11 months."</td> </tr> <tr> <td>Male Workforce Participation Rate</td> <td>Annual workforce participation rate for males.</td> </tr> <tr> <td>Female Workforce Participation Rate</td> <td>Annual workforce participation rate for females.</td> </tr> </tbody> </table>	Column	Description	Year	The calendar year for each value	Annual Average Workforce Participation	Defined by the Bureau of Labor Statistics as "the percentage of the population [16 years and older] that is either employed or unemployed (that is, either working or actively seeking work). Note that 2015's Annual Average is calculated using the first 11 months."	Male Workforce Participation Rate	Annual workforce participation rate for males.	Female Workforce Participation Rate	Annual workforce participation rate for females.	Apr 3, 2017	10KB	2
Column	Description												
Year	The calendar year for each value												
Annual Average Workforce Participation	Defined by the Bureau of Labor Statistics as "the percentage of the population [16 years and older] that is either employed or unemployed (that is, either working or actively seeking work). Note that 2015's Annual Average is calculated using the first 11 months."												
Male Workforce Participation Rate	Annual workforce participation rate for males.												
Female Workforce Participation Rate	Annual workforce participation rate for females.												

Result History Upgrade

StatCrunch keeps a live result history found at the following menu location: *StatCrunch > Results*. This list keeps track of all statistical and graphical results that are created within the current StatCrunch session, with the ability to show previously closed results.

The result history has been upgraded to include more informative titles that can help distinguish between multiple results from the same feature. Titles now display information about which data columns were used within the result. Depending on the feature, other key pieces of information are included within the title. These result titles will additionally include hover information specifying any *Where Statement* or *Group by* variable.

Below is an example using the [Movie Budgets and Box Office Earnings](#) data set. Three statistical results shown below come from the same statistical feature: *Stat > Summary Stats > Columns*. The titles for each summary stats feature shows the columns of data used by the feature while additional *Where* and *Group by* details appear by hovering over the result history.

Movie Budgets and Box Office Earnings (Updated Fall 2016)

StatCrunch | Applets | Edit | Data | Stat | Graph | Help

Month | Day | Release Year | Budget(\$M) | Domestic Gr | Worldwide G | var8

8 | Tangled | Nov | 24 | 2010 | 260 | 200.82194 | 586.58194

9 | Spider-Man | May | 4 | 2007 | 258 | 336.5303 | 890.8753

10 | Avengers: A | May | 1 | 2015 | 250 | 459.00587 | 1404.7059

11 | Captain Ame | May | 6 | 2016 | 250 | 408.08435 | 1151.6843

12 | Batman v Si | Mar | 25 | 2016 | 250 | 330.36019 | 868.16019

13 | The Hobbit: | Dec | 14 | 2012 | 250 | 303.00357 | 1017.0036

14 | Harry Potter | Jul | 15 | 2009 | 250 | 301.9592 | 935.08369

15 | The Hobbit: | Dec | 13 | 2013 | 250 | 258.36685 | 960.36685

16 | Transformer | Jun | 2 | 2009 | 250 | 301.9592 | 935.08369

17 | Transformer | Jun | 2 | 2009 | 250 | 301.9592 | 935.08369

28 | Transformer | Jun | 27 | 2014 | 210 | 245.43908 | 1104.0391

29 | X-Men: The | May | 26 | 2006 | 210 | 234.36246 | 459.35956

30 | Robin Hood | May | 14 | 2010 | 210 | 105.26073 | 322.24150

Result #1
Summary statistics:
Column | n | Mean | Variance | Std. dev. | Std. err.
Budget(\$M) | 5222 | 30.678772 | 1586.0874 | 39.825713 | 0.5511867

Result #2
Summary statistics:
Where: "Release Year" > 1990
Column | n | Mean | Variance | Std. dev. | Std.
Domestic Gross(\$M) | 4642 | 40.895199 | 4390.4645 | 66.26058 | 0.972

Result #3
Summary statistics for Worldwide Gross(\$M):
Group by: Month
Month | n | Mean | Variance | Std. dev. | Std. err. | Median | Range | Min | Max | Q1 | Q3
Jan | 339 | 38.428732 | 3412.4621 | 58.416283 | 3.172737 | 14.792779 | 518.62862 | 0 | 518.62862 | 2.712293 | 52.189039
Feb | 340 | 65.114883 | 9484.7663 | 97.389765 | 5.2817016 | 33.097835 | 783.77071 | 0 | 783.77071 | 5.345178 | 82.241316
Mar | 420 | 70.145068 | 15996.413 | 126.47693 | 6.171442 | 24.622252 | 1025.4911 | 0 | 1025.4911 | 3.644301 | 81.13923
Apr | 406 | 53.070017 | 13845.263 | 117.66589 | 5.8396603 | 21.103431 | 1514.0191 | 0 | 1514.0191 | 2.840417 | 62.063972

New Applet: Randomization Test for Slope

A new applet (**Applets > Resampling > Randomization test for slope**) has been added that performs a randomization test for a slope between two quantitative variables. The applet operates very similarly to the current randomization test for correlation applet, but tracks slope instead of the linear correlation coefficient.

Below is an example of the new applet using the [Fat and calorie content for a sample of seven chicken sandwiches](#) data set. The applet below shows 10,000 randomized simulations where the values for the *Calories* column are randomly assigned to values of the *Fat* column. The graph displays the slopes found for the 10,000 simulations with slopes highlighted in red that have a greater magnitude than the original slope between *Fat* and *Calories*.

