

# Data resources

Jeffrey Leek

May 18, 2016

# Open Government Sites

- ▶ United Nations <http://data.un.org/>
- ▶ U.S. <http://www.data.gov/>
- ▶ List of cities/states with open data
- ▶ United Kingdom <http://data.gov.uk/>
- ▶ France <http://www.data.gouv.fr/>
- ▶ Ghana <http://data.gov.gh/>
- ▶ Australia <http://data.gov.au/>
- ▶ Germany <https://www.govdata.de/>
- ▶ Hong Kong <http://www.gov.hk/en/theme/psi/datasets/>
- ▶ Japan <http://www.data.go.jp/>
- ▶ Many more <http://www.data.gov/opendatasites>

Gapminder for a fact-based world view

Blog | FAQ | About | Contact | Donate

Search this site...

HOME | GAPMINDER WORLD | **DATA** | VIDEOS | DOWNLOADS | FOR TEACHERS | LABS

Browse: Home / Data

## Data in Gapminder World

List of indicators [About countries & territories](#) [Documentation](#) [Data blog](#)

The table below lists all indicators displayed in Gapminder World. Click the name of the indicator or the data provider to access information about the indicator and a link to the data provider. Indicators labeled "Various sources" are compiled by Gapminder. They can be reused freely but please attribute Gapminder.

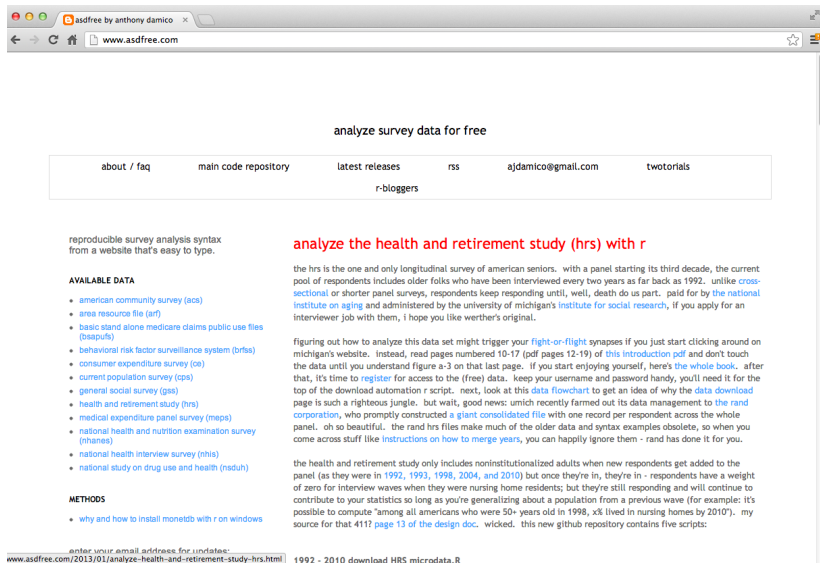
### List of indicators in Gapminder World

Show 25 Indicators Search:

Indicator name	Data provider	Category	Subcategory	Download	View	Visualize
Adults with HIV (%), age 15-49	Based on UNAIDS	Health	HIV			
Age at 1st marriage (women)	Various sources	Population				
Aged 15+ employment rate (%)	International Labour Organization	Work	Employment rate			
Aged 15+ labour force participation rate (%)	International Labour Organization	Work	Labour force participation			
Aged 15+ unemployment rate (%)	International Labour Organization	Work	Unemployment			
Aged 15-24 employment rate (%)	International Labour Organization	Work	Employment rate			

<http://www.gapminder.org/>

# Survey data from the United States



analyze survey data for free

about / faq    main code repository    latest releases    rss    ajdamico@gmail.com    twotorials

r-bloggers

reproducible survey analysis syntax  
from a website that's easy to type.

**AVAILABLE DATA**

- american community survey (acs)
- area resource file (arf)
- basic stand alone medicare claims public use files (bsapufs)
- behavioral risk factor surveillance system (brfss)
- consumer expenditure survey (ce)
- current population survey (cps)
- general social survey (gss)
- health and retirement study (hrs)
- medical expenditure panel survey (meps)
- national health and nutrition examination survey (nhanes)
- national health interview survey (nhis)
- national study on drug use and health (nsduh)

**METHODS**

- why and how to install monetdb with r on windows

enter your email address for updates:  
[www.asdfree.com/2013/01/analyze-health-and-retirement-study-hrs.html](http://www.asdfree.com/2013/01/analyze-health-and-retirement-study-hrs.html)

1992 - 2010 download HRS microdata.R

**analyze the health and retirement study (hrs) with r**

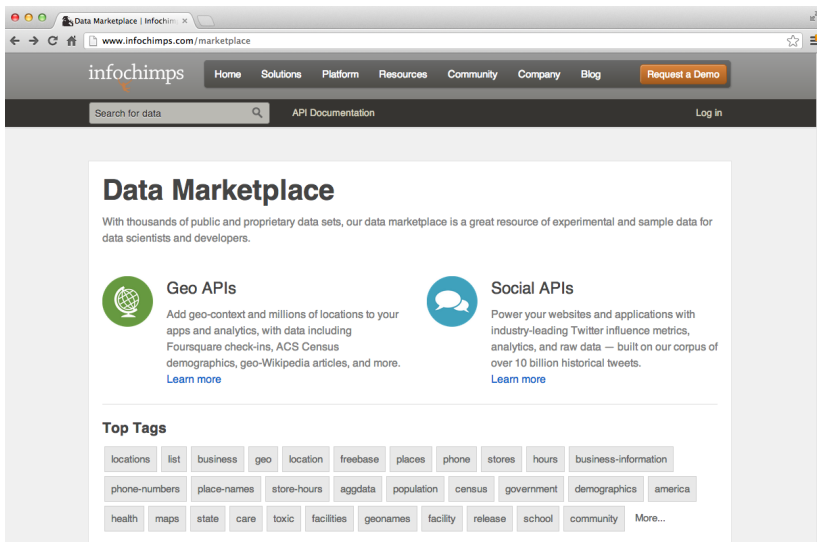
the hrs is the one and only longitudinal survey of american seniors. with a panel starting its third decade, the current pool of respondents includes older folks who have been interviewed every two years as far back as 1992. unlike [cross-sectional](#) or shorter panel surveys, respondents keep responding until, well, death do us part. paid for by the [national institute on aging](#) and administered by the university of michigan's [institute for social research](#), if you apply for an interviewer job with them, i hope you like werther's original.

figuring out how to analyze this data set might trigger your *fight-or-flight* synapses if you just start clicking around on michigan's website. instead, read pages numbered 10-17 (pdf pages 12-19) of [this introduction pdf](#) and don't touch the data until you understand figure a-3 on that last page. if you start enjoying yourself, here's [the whole book](#). after that, it's time to [register](#) for access to the (free) data. keep your username and password handy, you'll need it for the top of the download automation r script. next, look at this [data flowchart](#) to get an idea of why the [data download](#) page is such a righteous jungle. but wait, good news: umich recently farmed out its data management to [the rand corporation](#), who promptly constructed a [giant consolidated file](#) with one record per respondent across the whole panel. oh so beautiful. the rand hrs files make much of the older data and syntax examples obsolete, so when you come across stuff like [instructions on how to merge years](#), you can happily ignore them - rand has done it for you.

the health and retirement study only includes noninstitutionalized adults when new respondents get added to the panel (as they were in 1992, 1993, 1998, 2004, and 2010) but once they're in, they're in - respondents have a weight of zero for interview waves when they were nursing home residents; but they're still responding and will continue to contribute to your statistics so long as you're generalizing about a population from a previous wave (for example: it's possible to compute "among all americans who were 50+ years old in 1998, x% lived in nursing homes by 2010"). my source for that 4117 [page 13 of the design doc](#). wicked. this new github repository contains five scripts:

<http://www.asdfree.com/>

# Infochimps Marketplace




The screenshot shows the Infochimps Marketplace website in a web browser. The browser's address bar displays `www.infochimps.com/marketplace`. The website's navigation bar includes links for Home, Solutions, Platform, Resources, Community, Company, and Blog, along with a 'Request a Demo' button. A search bar with the placeholder text 'Search for data' and a magnifying glass icon is positioned on the left, while 'API Documentation' and a 'Log in' link are on the right. The main content area features a large heading 'Data Marketplace' followed by a descriptive paragraph. Below this, there are two featured sections: 'Geo APIs' with a globe icon and 'Social APIs' with a speech bubble icon. Each section provides a brief description of the data available and a 'Learn more' link. At the bottom, a 'Top Tags' section displays a grid of tags such as 'locations', 'list', 'business', 'geo', 'location', 'freebase', 'places', 'phone', 'stores', 'hours', 'business-information', 'phone-numbers', 'place-names', 'store-hours', 'aggdata', 'population', 'census', 'government', 'demographics', 'america', 'health', 'maps', 'state', 'care', 'toxic', 'facilities', 'geonames', 'facility', 'release', 'school', 'community', and a 'More...' link.

www.infochimps.com/marketplace


## Data Marketplace

With thousands of public and proprietary data sets, our data marketplace is a great resource of experimental and sample data for data scientists and developers.



### Geo APIs

Add geo-context and millions of locations to your apps and analytics, with data including Foursquare check-ins, ACS Census demographics, geo-Wikipedia articles, and more.  
[Learn more](#)



### Social APIs

Power your websites and applications with industry-leading Twitter influence metrics, analytics, and raw data — built on our corpus of over 10 billion historical tweets.  
[Learn more](#)

### Top Tags

locations	list	business	geo	location	freebase	places	phone	stores	hours	business-information
phone-numbers	place-names	store-hours	aggdata	population	census	government	demographics	america		
health	maps	state	care	toxic	facilities	geonames	facility	release	school	community
										More...

`http://www.infochimps.com/marketplace`

# What's in your data?

## Participate in competitions

Kaggle is an arena where you can match your data science skills against a global cadre of experts in statistics, mathematics, and machine learning. Whether you're a world-class algorithm wizard competing for prize money or a novice looking to learn from the best, here's your chance to jump in and geek out, for fame, fortune, or fun.

**Join as a participant**

(Need convincing?)

## Create a competition

Kaggle is a platform for data prediction competitions that allows organizations to post their data and have it scrutinized by the world's best data scientists. In exchange for a prize, winning competitors provide the algorithms that beat all other methods of solving a data crunching problem. Most data problems can be framed as a competition.

**Learn more about hosting**

`http://www.kaggle.com/`

# Collections by data scientists

- ▶ Hilary Mason <http://bitly.com/bundles/hmason/1>
- ▶ Peter Skomoroch  
<https://delicious.com/pskomoroch/dataset>
- ▶ Jeff Hammerbacher  
<http://www.quora.com/Jeff-Hammerbacher/Introduction-to-Data-Science-Data-Sets>
- ▶ Gregory Piatetsky-Shapiro  
<http://www.kdnuggets.com/gps.html>
- ▶ <http://blog.mortardata.com/post/67652898761/6-dataset-lists-curated-by-data-scientists>

# More specialized collections

- ▶ Stanford Large Network Data
- ▶ UCI Machine Learning
- ▶ KDD Nugets Datasets
- ▶ CMU Statlib
- ▶ Gene expression omnibus
- ▶ ArXiv Data
- ▶ Public Data Sets on Amazon Web Services



# Some API's with R interfaces

- ▶ twitter and twitteR package
- ▶ figshare and rfigshare
- ▶ PLoS and rplos
- ▶ rOpenSci
- ▶ Facebook and RFacebook
- ▶ Google maps and RGoogleMaps