

OCT 13, 2015

19 Free Public Data Sets For Your First Data Science Project

👤 Rajit Dasgupta 💬 2 📌 DATA SCIENCE ([HTTPS://WWW.SPRINGBOARD.COM/BLOG/CATEGORY/DATA-SCIENCE/](https://www.springboard.com/blog/category/data-science/)), LISTS ([HTTPS://WWW.SPRINGBOARD.COM/BLOG/CATEGORY/LISTS/](https://www.springboard.com/blog/category/lists/))



Completing your first project is a major milestone on the road to becoming a data scientist. It's also an intimidating process. The first step is to find an appropriate, interesting data set. You should decide how large and how messy a dataset you want to work with; while cleaning data is an integral part of data science, you may want to start with clean dataset for your first project so that you can focus on the analysis rather than on cleaning the data.

Based on the learnings from our [Foundations of Data Science Workshop](https://www.springboard.com/workshops/data-science) (<https://www.springboard.com/workshops/data-science>) and the [Data Science Career Track](https://www.springboard.com/workshops/data-science-career-track/) (<https://www.springboard.com/workshops/data-science-career-track/>), we've selected datasets of varying types and complexity that we think work well for first projects (some of them work for research projects as well!). These data-sets cover a variety of sources: demographic data, economic data, text data, and corporate data.

1. **United States Census Data:** The United States Census publishes reams of

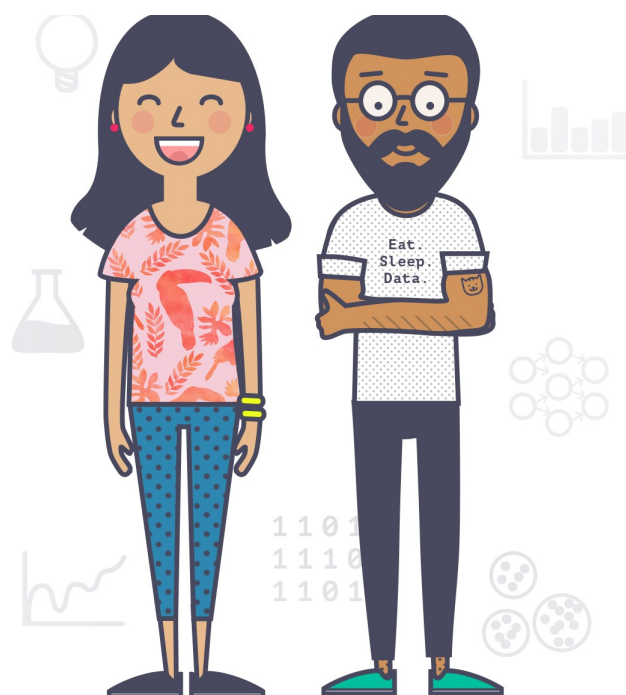
demographic data at the state, city, and even zip code level. The data set is fantastic for creating geographic data visualizations and can be accessed on the Census Website (http://www2.census.gov/acs2013_1yr/summaryfile/). Alternatively, the data can be accessed via an API. One convenient way to use that API is through the `chloroplethr` (<https://cran.r-project.org/web/packages/choroplethr/>). In general, this data is very clean and very comprehensive.

2. **FBI Crime Data:** The FBI crime data set is fascinating. If you're interested in analyzing time series data, you can use it to chart changes in crime rates at the national level over a 20 year period (https://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2013/crime-in-the-u.s.-2013/tables/1tabledatadecoverviewpdf/table_1_crime_in_the_united_states_by_volume_and_rate_per_100000_inhabitants_1994-2013). Alternatively, you can look at the data geographically (https://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2013/crime-in-the-u.s.-2013/tables/table-8/table_8_offenses_known_to_law_enforcement_by_state_by_city_2013.xls/view).
3. **CDC Cause of Death:** The Center for Disease Control control maintains a database on cause of death (<http://wonder.cdc.gov/>). The data can be segmented in almost every way imaginable: age, race, year, and so on.
4. **Medicare Hospital Quality:** Medicare maintains a database on complication rates by hospital (<https://data.medicare.gov/data/hospital-compare#>) that provides for interesting comparisons.
5. **SEER Cancer Incidence:** The US government also has data about cancer incidence (<http://seer.cancer.gov/faststats/selections.php?series=cancer>), again segmented by age, race, gender, year, and other factors.
6. **Bureau of Labor Statistics:** Many important economic indicators for the United States (like unemployment and inflation) can be found on the Bureau of Labor Statistics website (<http://www.bls.gov/data/>). Most of the data can be segmented both by time and by geography.

7. **The Bureau of Economic Analysis:** The Bureau of Economic Analysis (<http://www.bea.gov/national/index.htm>) also has national and regional economic data, like GDP and exchange rates.
8. **IMF Economic Data:** If you want a view of international data, you can find it on the IMF website (<http://data.imf.org/>).
9. **Dow Jones Weekly Returns:** Predicting stock prices is a major application of data analysis and machine learning. One dataset to explore is the weekly returns of the Dow Jones Index. (<http://archive.ics.uci.edu/ml/datasets/Dow+Jones+Index>)
10. **Boston Housing Data:** The Boston Housing Data Set (<http://archive.ics.uci.edu/ml/datasets/Housing>) contains median housing prices in Boston suburbs as well as 13 attributes that contribute to those prices. It's an excellent set for experimenting with various types of regressions.
11. **Enron Emails:** After the collapse of Enron, a dataset of roughly 500,000 emails with message text and metadata were released. The dataset (<http://www.cs.cmu.edu/~enron/>) is now famous and provides an excellent testing ground for text related analysis. It has the messiness of real world data.
12. **Google N-Grams:** If you're interested in truly massive data, the Google n-grams (<http://aws.amazon.com/datasets/8172056142375670>) dataset counts the frequency of words and phrases by year across a huge number of text sources. The resulting file is 2.2 TB.
13. **Sentence Sentiments:** Researchers have labeled 3,000 sentences (<https://archive.ics.uci.edu/ml/datasets/Sentiment+Labelled+Sentences>) as expressing positive or negative sentiments. If you're interested in classifying text, this is a great place to start.

14. **Reddit Comments:** Reddit released a dataset of every comment that has ever been made on the site. That's over a terabyte of data uncompressed, so if you want a smaller dataset to work with Kaggle has hosted the comments from May 2015 (<https://www.kaggle.com/c/reddit-comments-may-2015/>) on their site.
15. **Wikipedia:** Wikipedia provides instructions for downloading the text of English language articles (https://en.wikipedia.org/wiki/Wikipedia:Database_download#English-language_Wikipedia).
16. **Lending Club:** Lending Club (<https://www.lendingclub.com/info/download-data.action>) provides data about loan applications it has rejected as well as the performance of loans that it issued. The dataset lends itself both to categorization techniques (will a given loan default) as well as regressions (how much will be paid back on a given loan.)
17. **Walmart:** Walmart has released store level sales data (<https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/data>) for 98 items across 45 stores. This is an excellent data for time series analysis and has interesting seasonal components as well.
18. **Airbnb:** [This website](http://tomslee.net/airbnb-data-collection-get-the-data) (<http://tomslee.net/airbnb-data-collection-get-the-data>) offers different datasets related to Airbnb and listings related to different cities.
19. **Yelp:** Yelp releases an academic dataset (https://www.yelp.com/academic_dataset) that contains information for the areas around 30 universities.

Well – now it's time to get cracking! If you want to jumpstart your Data Science career today, I'd recommend checking out our [12-Week Online Workshop – Foundations of Data Science](https://www.springboard.com/workshops/data-science) (<https://www.springboard.com/workshops/data-science>). Head [here](https://www.springboard.com/workshops/data-science) (<https://www.springboard.com/workshops/data-science>) for more on that. If you wanted even more resources, check out the [Springboard](https://www.springboard.com) (<https://www.springboard.com>) home page.



Springboard

Get mentored
1-to-1 by top
data scientists

Get a Data
Science job,
Guaranteed!

www.springboard.com

(https://www.springboard.com/workshops/data-science-career-track?utm_source=blog&utm_campaign=springboardbottombanner&utm_medium=blog)

Data Science Career Track

LEARN UX DESIGN WITH SPRINGBOARD



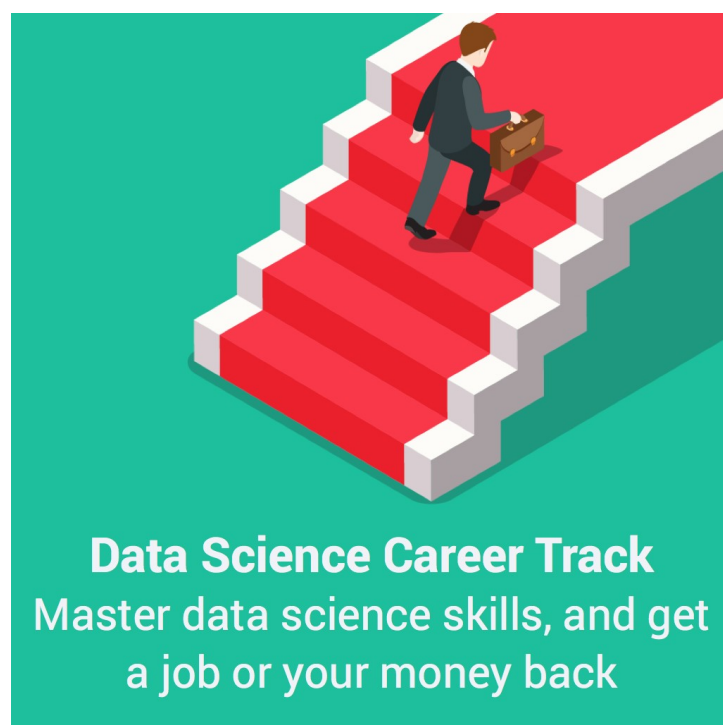
**LEARN
UX DESIGN**

In 8 Weeks, with a Mentor.

GET ACCESS >>

(https://www.springboard.com/workshops/ux-design?utm_source=blog&utm_medium=uxsidebar&utm_campaign=uxblogsidebar)

LEARN DATA SCIENCE WITH SPRINGBOARD



Data Science Career Track
Master data science skills, and get
a job or your money back

(<https://www.springboard.com/workshops/data->

science-career-track?utm_source=blog&
utm_medium=dssidebar&
utm_campaign=dsblogsidebar)

LEARN DATA ANALYSIS WITH SPRINGBOARD

Related Articles

Weekly MOOC Buffet 13 – 14 Free
Courses You Can Start This Week!
([https://www.springboard.com
/blog/weekly-mooc-buffet-13-14-
free-courses-you-can-start-this-week/](https://www.springboard.com/blog/weekly-mooc-buffet-13-14-free-courses-you-can-start-this-week/))



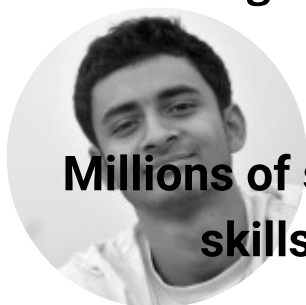
Weekly MOOC Buffet 27 (6th Oct – 12th
Oct, 2014)
([https://www.springboard.com
/blog/weekly-mooc-buffet-27/](https://www.springboard.com/blog/weekly-mooc-buffet-27/))

([https://www.springboard.com/workshops
/analytics?utm_source=blog&
utm_medium=dasidebar&
utm_campaign=dablogsidebar](https://www.springboard.com/workshops/analytics?utm_source=blog&utm_medium=dasidebar&utm_campaign=dablogsidebar))

AUTHOR

Rajit Dasgupta

([https://www.springboard.com
/blog/author/rajit/](https://www.springboard.com/blog/author/rajit/))



(<https://www.springboard.com>)
**Millions of students are improving their
skills by using our courses**

</blog/author/rajit/>)

Growth Lead , Springboard ()

LEARN MORE

(</workshops/>)

Stay in the know!

Workshops



I Want To Learn... ▼

Resources ▼

Company ▼

Enter your email **GET UPDATES**

(https://twitter.com/springboard)
(https://facebook.com/springboard)