# Downloading Data from World Bank and US Government Websites

To explore the relationship between economic factors and birth rate, locate and download economic indicator and birth rate data.

## Project Overview

### Data

Data on gross domestic product and birth rate for the world, the United States of America, and five groups of countries aggregated by income was located on the World Bank Data Bank website. Data for the period 1998-2015 was downloaded in comma-separated value (CSV) format. Information on birth rate by US State was obtained from the Centers for Disease Control’s WONDER engine in text format for the period 2003-2015. Number of births by US State was available for the period 1995-2002. State population estimates were downloaded from the US Census Bureau’s site in a variety of formats, as was *per capita* income for the period 1967-2015.

### Technologies

Files were extracted from the compressed downloads using WinZip. Unzipped files were examined using either a text editor or Microsoft Excel.

### Techniques

Headers for the text files were edited to remove whitespace. Excel and CSV files were checked for extraneous information and edited to create software-friendly headers before being saved in CSV format. Small datasets in tabular format were separated into different tables and reformed into time series format.

### Software

Data was downloaded to a desktop computer running Windows. Notepad and Excel were used to edit data headers.

### Lessons Learned

Three separate websites were accessed to obtain required data. After decompression, data was stored in three different formats. Data coverage varied by year, with data from more recent periods being more complete.

### Applications to Business

Governments and non-governmental organizations (NGOs) such as the World Bank make a wealth of information available for public use, but it is not consistently formatted. Organizations that provide for tailored download of data products include the CDC and the World Bank.

## Table of Contents

### [Project Background](#Background)

### [Step 1 – Locating the data](#Step1)

### [Step 2 – Exporting the data](#Step2)

### [Step 3 – Cleaning the data](#Step3)

### [Conclusion](#Conclusion)

### [Related Projects](#Related)

### [Bio](#Bio)

### [Contact Information](#Contact)

## Project Background

A Pew Research Center report published April 6, 2010 linked a decline in the US birth rate to the recession of 2008. They used birth data from 25 US States for which 2008 full-year records were available. This project was intended to obtain data that could be used to investigate whether the trend was localized to the US and to see if the birth rate increased after the initial drop.

## Step 1 – Locating the data

Prior knowledge indicated that the World Bank’s website would offer economic indicator information for many countries. A web search for US birth rate led to the CDC website and WONDER. A lack of birth rate data from that site occasioned a second search for US State populations by year, which data was found on the US Census Bureau site. On that site, additional browsing uncovered data on *per capita* income for several years.

## Step 2 – Exporting the data

The World Bank offers a download service providing CSV files with data and explanations stored separately. For this project, three measures were selected: Birth rate, crude (per 1,000 people); Gross Domestic Product (GDP) per capita (current US$); and GDP per capita, adjusted for purchasing power parity (PPP) (current international $). These series were gathered for the period from 1998 through 2015 for seven regions: high income countries; low income countries; lower middle income countries; middle income countries; upper middle income countries; the United States of America; and the world. Data was downloaded from <http://databank.worldbank.org/data/reports.aspx?source=world-development-indicators>, and had last been updated 17-APR-2017.

WONDER allows users to choose which measures are to be included and how they are to be grouped. The chosen data may then be stored in a text file. Data on natality were downloaded from <https://wonder.cdc.gov/natality-v2002.html> for the years 1995-2002, <https://wonder.cdc.gov/natality-v2006.html> for the years 2003-2006, and <https://wonder.cdc.gov/natality-current.html> for the years 2007-2015.

Census data is available in several different formats depending on the age of the information. Excel and CSV formats are most common. State population projections based on 1990 Census data were downloaded from <https://www.census.gov/population/projections/data/state/st_yr95to00.html> for the years 1995 through 2000 and from <https://www.census.gov/population/projections/data/state/st_yr01to05.html> for the years 2001 through 2005. Information on per capita income was downloaded from <https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-income-people.html> for the period 1967 through 2015.

## Step 3 – Cleaning the data

Downloaded compressed files were unzipped and examined using either Excel or Notepad. World Bank files contained citation information and whitespace in the header text. Both were removed, and the headers replaced with contiguous text. CDC text files also contained whitespace in some column headers, which was removed. Census data was structured differently depending on the file. In general, headers were regularized and citation information removed.

## Conclusion

Data is widely available on the web, but it must be discovered. Downloaded information is not generally usable as-is, but small data files can be made ready for use by editing in word processing and spreadsheet programs.

## Related Projects

Visualizing US Birth Rate from 1998 to 2015

Visualizing Birth Rates and Economic Indicators from 1998 to 2015

## Bio

Kimberley Elcess is a data scientist with a background in materials science and engineering who focuses on solving problems in capital-intensive industries.

Specifically, Kimberley employs statistics, machine learning, and visualization to solve problems and generate insight in entrepreneurial environments using Excel, R, and Bayesian inference.

## Contact Information

Kimberley Elcess

[kelcess@gmail.com](mailto:kelcess@gmail.com)

[www.kimberleyelcess.com](http://www.kimberleyelcess.com)