Data Analysis in Linux

CJ Fearnley
cjf@LinuxForce.net
http://www.LinuxForce.net
cjf@CJFearnley.com
http://blog.CJFearnley.com

3 July 2013

Presentation to ***PLUG: Philadelphia area Linux Users' Group*** at the University of the Sciences (USP) in Philadelphia

On-line version of this presentation:
http://www.CJFearnley.com/Data.Analysis.PLUG.July2013.pdf

My Background

- I earned a BA in Mathematical Sciences and Philosophy from Binghamton University in 1989. But I skipped statistics.
- Some years ago I read the 2004 edition of Nassim Nicholas Taleb's *Fooled by Randomness: The Hidden Role of Chance in live and in the Markets* which was fascinating but excessively irreverent.
- In 2011, I read Leonard Mlodinow's 2008 bestseller *The Drunkard's Walk: How Randomness Rules Our Lives* which was exquisite.

My Background

- I earned a BA in Mathematical Sciences and Philosophy from Binghamton University in 1989. But I skipped statistics.
- Some years ago I read the 2004 edition of Nassim Nicholas Taleb's *Fooled by Randomness: The Hidden Role of Chance in live and in the Markets* which was fascinating but excessively irreverent.
- In 2011, I read Leonard Mlodinow's 2008 bestseller *The Drunkard's Walk: How Randomness Rules Our Lives* which was exquisite.

My Background

- I earned a BA in Mathematical Sciences and Philosophy from Binghamton University in 1989. But I skipped statistics.
- Some years ago I read the 2004 edition of Nassim Nicholas Taleb's *Fooled by Randomness: The Hidden Role of Chance in live and in the Markets* which was fascinating but excessively irreverent.
- In 2011, I read Leonard Mlodinow's 2008 bestseller *The Drunkard's Walk: How Randomness Rules Our Lives* which was exquisite.

My Background continued

- That led to me to write the essays *Are Randomness and Uncertainty fundamental and pervasive?*

  `http://blog.cjfearnley.com/2011/04/20/are-randomness-and-uncertainty-fundamental-and-p`

  and *Determinism and Randomness Always and Only Coexist*

  `http://blog.cjfearnley.com/2012/02/21/determinism-and-randomness-always-and-only-coexi`

- Then last summer I read David Salsburg's *The Lady Tasting Tea* which was only good, but it taught me that there is no theory of statistics. Statistics is, at present, just a bunch of mathematical tricks which have been collected over time to try to understand something which we do not understand. Yet statistics is the basis of all modern science and as such it is the basis of knowledge and of modern civilization.

My Background continued

- That led to me to write the essays *Are Randomness and Uncertainty fundamental and pervasive?*

  `http://blog.cjfearnley.com/2011/04/20/are-randomness-and-uncertainty-fundamental-and-p`

  and *Determinism and Randomness Always and Only Coexist*

  `http://blog.cjfearnley.com/2012/02/21/determinism-and-randomness-always-and-only-coexi`

- Then last summer I read David Salsburg's *The Lady Tasting Tea* which was only good, but it taught me that there is no theory of statistics. Statistics is, at present, just a bunch of mathematical tricks which have been collected over time to try to understand something which we do not understand. Yet statistics is the basis of all modern science and as such it is the basis of knowledge and of modern civilization.

Computing for Data Analysis

- Coursera course *Computing for Data Analysis* with Roger Peng of Johns Hopkins University
- `http://www.coursera.org/course/compdata`
- Intro video: `https://www.youtube.com/watch?v=gk6E57H6mTs`
- Next offering in September (4 weeks)
- Computing for Data Analysis: Week 1

  `http://www.youtube.com/playlist?list=PLjTlxb-wKvXNSDfcKPFH2gzHGyjpeCZmJ`

  Computing for Data Analysis: Week 2

  `http://www.youtube.com/playlist?list=PLjTlxb-wKvXNnjUTX4C8IeIhPBjPkng6B`

  Computing for Data Analysis: Week 3

  `http://www.youtube.com/playlist?list=PLjTlxb-wKvXOzI2h0F2_rYZHIXz8GWBop`

  Computing for Data Analysis: Week 4

  `http://www.youtube.com/playlist?list=PLjTlxb-wKvXOdzysAE6qrEBN_aSBC0LZS`

Computing for Data Analysis

- Coursera course *Computing for Data Analysis* with Roger Peng of Johns Hopkins University
- `http://www.coursera.org/course/compdata`
- Intro video: `https://www.youtube.com/watch?v=gk6E57H6mTs`
- Next offering in September (4 weeks)
- Computing for Data Analysis: Week 1

  `http://www.youtube.com/playlist?list=PLjTlxb-wKvXNSDfcKPFH2gzHGyjpeCZmJ`

  Computing for Data Analysis: Week 2

  `http://www.youtube.com/playlist?list=PLjTlxb-wKvXNnjUTX4C8IeIhPBjPkng6B`

  Computing for Data Analysis: Week 3

  `http://www.youtube.com/playlist?list=PLjTlxb-wKvXOzI2h0F2_rYZHIXz8GWBop`

  Computing for Data Analysis: Week 4

  `http://www.youtube.com/playlist?list=PLjTlxb-wKvXOdzysAE6qrEBN_aSBC0LZS`

Computing for Data Analysis

- Coursera course *Computing for Data Analysis* with Roger Peng of Johns Hopkins University
- `http://www.coursera.org/course/compdata`
- Intro video: `https://www.youtube.com/watch?v=gk6E57H6mTs`
- Next offering in September (4 weeks)
- Computing for Data Analysis: Week 1

  `http://www.youtube.com/playlist?list=PLjTlxb-wKvXNSDfcKPFH2gzHGyjpeCZmJ`

  Computing for Data Analysis: Week 2

  `http://www.youtube.com/playlist?list=PLjTlxb-wKvXNnjUTX4C8IeIhPBjPkng6B`

  Computing for Data Analysis: Week 3

  `http://www.youtube.com/playlist?list=PLjTlxb-wKvXOzI2h0F2_rYZHIXz8GWBop`

  Computing for Data Analysis: Week 4

  `http://www.youtube.com/playlist?list=PLjTlxb-wKvXOdzysAE6qrEBN_aSBC0LZS`

CapeTown Open Education Declaration

"We are on the cusp of a global revolution in teaching and learning. Educators worldwide are developing a vast pool of educational resources on the Internet, open and free for all to use. These educators are creating a world where each and every person on earth can access and contribute to the sum of all human knowledge. They are also planting the seeds of a new pedagogy where educators and learners create, shape and evolve knowledge together, deepening their skills and understanding as they go."

CapeTown Open Education Declaration, 2007

`http://www.capetowndeclaration.org/read-the-declaration`

See my presentation last year on "Education Automation Now and in the Future"

`http://www.cjfearnley.com/Asheville.Education.Autom`

CapeTown Open Education Declaration

"We are on the cusp of a global revolution in teaching and learning. Educators worldwide are developing a vast pool of educational resources on the Internet, open and free for all to use. These educators are creating a world where each and every person on earth can access and contribute to the sum of all human knowledge. They are also planting the seeds of a new pedagogy where educators and learners create, shape and evolve knowledge together, deepening their skills and understanding as they go."

CapeTown Open Education Declaration, 2007

`http://www.capetowndeclaration.org/read-the-declaration`

See my presentation last year on "Education Automation Now and in the Future"

`http://www.cjfearnley.com/Asheville.Education.Autom`

Data Analysis

- Coursera course *Data Analysis* with Jeff Leek of Johns Hopkins University
- `http://www.coursera.org/course/dataanalysis`
- Intro video: `http://www.youtube.com/watch?v=-lutj1vrPwQ`
- Next offering in January (8 weeks)
- YouTube videos: `http://www.youtube.com/user/jtleek2007`

Data Analysis

- Coursera course *Data Analysis* with Jeff Leek of Johns Hopkins University
- `http://www.coursera.org/course/dataanalysis`
- Intro video: `http://www.youtube.com/watch?v=-lutj1vrPwQ`
- Next offering in January (8 weeks)
- YouTube videos: `http://www.youtube.com/user/jtleek2007`

Data Analysis

- Coursera course *Data Analysis* with Jeff Leek of Johns Hopkins University
- `http://www.coursera.org/course/dataanalysis`
- Intro video: `http://www.youtube.com/watch?v=-lutj1vrPwQ`
- Next offering in January (8 weeks)
- YouTube videos: `http://www.youtube.com/user/jtleek2007`

Steps in a data analysis

- Define the question
- Define the ideal data set
- Determine what data you can access
- Obtain the data
- Clean the data
- Exploratory data analysis
- Statistical prediction/modeling
- Interpret results
- Challenge results
- Synthesize/write up results
- Create reproducible code

Steps in a data analysis

- Define the question
- Define the ideal data set
- Determine what data you can access
- Obtain the data
- Clean the data
- Exploratory data analysis
- Statistical prediction/modeling
- Interpret results
- Challenge results
- Synthesize/write up results
- Create reproducible code

Steps in a data analysis

- Define the question
- Define the ideal data set
- Determine what data you can access
- Obtain the data
- Clean the data
- Exploratory data analysis
- Statistical prediction/modeling
- Interpret results
- Challenge results
- Synthesize/write up results
- Create reproducible code

Steps in a data analysis

- Define the question
- Define the ideal data set
- Determine what data you can access
- Obtain the data
- Clean the data
- Exploratory data analysis
- Statistical prediction/modeling
- Interpret results
- Challenge results
- Synthesize/write up results
- Create reproducible code

Steps in a data analysis

- Define the question
- Define the ideal data set
- Determine what data you can access
- Obtain the data
- Clean the data
- Exploratory data analysis
- Statistical prediction/modeling
- Interpret results
- Challenge results
- Synthesize/write up results
- Create reproducible code

Steps in a data analysis

- Define the question
- Define the ideal data set
- Determine what data you can access
- Obtain the data
- Clean the data
- Exploratory data analysis
- Statistical prediction/modeling
- Interpret results
- Challenge results
- Synthesize/write up results
- Create reproducible code

Steps in a data analysis

- Define the question
- Define the ideal data set
- Determine what data you can access
- Obtain the data
- Clean the data
- Exploratory data analysis
- Statistical prediction/modeling
- Interpret results
- Challenge results
- Synthesize/write up results
- Create reproducible code

Steps in a data analysis

- Define the question
- Define the ideal data set
- Determine what data you can access
- Obtain the data
- Clean the data
- Exploratory data analysis
- Statistical prediction/modeling
- Interpret results
- Challenge results
- Synthesize/write up results
- Create reproducible code

Steps in a data analysis

- Define the question
- Define the ideal data set
- Determine what data you can access
- Obtain the data
- Clean the data
- Exploratory data analysis
- Statistical prediction/modeling
- Interpret results
- Challenge results
- Synthesize/write up results
- Create reproducible code

Steps in a data analysis

- Define the question
- Define the ideal data set
- Determine what data you can access
- Obtain the data
- Clean the data
- Exploratory data analysis
- Statistical prediction/modeling
- Interpret results
- Challenge results
- Synthesize/write up results
- Create reproducible code

Steps in a data analysis

- Define the question
- Define the ideal data set
- Determine what data you can access
- Obtain the data
- Clean the data
- Exploratory data analysis
- Statistical prediction/modeling
- Interpret results
- Challenge results
- Synthesize/write up results
- Create reproducible code

Background

- For an eight hour work day a human can produce 0.6 kWh (Kilowatt hours) of energy. So a human can produce 219 kWh per year.
- An ox can produce at a rate of 450 W (watts) or 3.6 kWh per day. So an ox can produce 1,314 kWh per year.

Background

- For an eight hour work day a human can produce 0.6 kWh (Kilowatt hours) of energy. So a human can produce 219 kWh per year.
- An ox can produce at a rate of 450 W (watts) or 3.6 kWh per day. So an ox can produce 1,314 kWh per year.

Define the Question

Source: Global Sustainable Energy: Past, Present and Future
`https://www.coursera.org/course/globalenergy`
What is the geographical location of the countries that consume less than 219 kWh per year? Note that these are the countries that are essentially living on the equivalent of human power as their source of energy.
Also, what is the geographical location of the countries that consume less than 1,314 kWh per year? Note that these are the countries that are essentially living on the equivalent of oxen power as their source of energy.
How many countries are there in this list?
In terms of population, what is the largest country in this list?

Define the Question

Source: Global Sustainable Energy: Past, Present and Future
`https://www.coursera.org/course/globalenergy`
What is the geographical location of the countries that consume less than 219 kWh per year? Note that these are the countries that are essentially living on the equivalent of human power as their source of energy.
Also, what is the geographical location of the countries that consume less than 1,314 kWh per year? Note that these are the countries that are essentially living on the equivalent of oxen power as their source of energy.
How many countries are there in this list?
In terms of population, what is the largest country in this list?

Define the Question

Source: Global Sustainable Energy: Past, Present and Future
`https://www.coursera.org/course/globalenergy`
What is the geographical location of the countries that consume less than 219 kWh per year? Note that these are the countries that are essentially living on the equivalent of human power as their source of energy.
Also, what is the geographical location of the countries that consume less than 1,314 kWh per year? Note that these are the countries that are essentially living on the equivalent of oxen power as their source of energy.
How many countries are there in this list?
In terms of population, what is the largest country in this list?

Define the Question

An Energy Slave is that quantity of energy (ability to do work) which, when used to construct and drive non-human infrastructure (machines, roads, power grids, fuel, draft animals, wind-driven pumps, etc.) replaces a unit of human labour (actual work). An energy slave does the work of a person, through the consumption of energy in the non-human infrastructure.
`http://en.wikipedia.org/wiki/Energy_Slave`

Define the ideal data set

Energy consumption data by country with population, energy consumption in kWh/y, and continent for columns.

Determine what data you can access

EIA: US Energy Information Administration

- International Energy Statistics, Total Primary Energy Consumption (Quadrillion Btu)

  `http://www.eia.gov/cfapps/ipdbproject/IEDIndex3.cfm?tid=44&pid=44&aid=2`

  Downloaded at 3 Apr 2013 at 06:05.
- International Energy Statistics, Population (Millions)

  `http://www.eia.gov/cfapps/ipdbproject/IEDIndex3.cfm?tid=93&pid=44&aid=33`

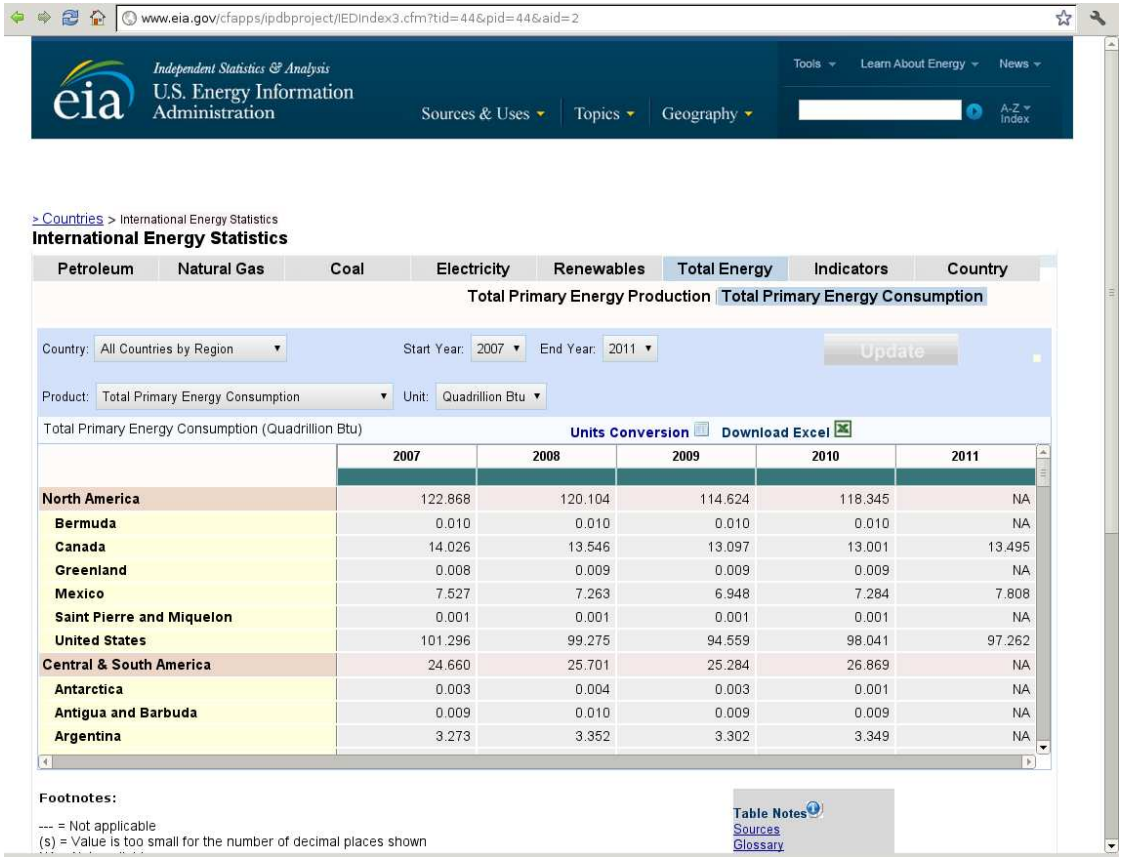
  Downloaded at 3 Apr 08:00.

Obtain the data 1

Go to the URLs. Click Download Excel.

International Energy Statistics, Total Primary Energy Consumption

`http://www.eia.gov/cfapps/ipdbproject/IEDIndex3.cfm?tid=44&pid=44&aid=2`



	2007	2008	2009	2010	2011
North America	122.868	120.104	114.624	118.345	NA
Bermuda	0.010	0.010	0.010	0.010	NA
Canada	14.026	13.546	13.097	13.001	13.495
Greenland	0.008	0.009	0.009	0.009	NA
Mexico	7.527	7.263	6.948	7.284	7.808
Saint Pierre and Miquelon	0.001	0.001	0.001	0.001	NA
United States	101.296	99.275	94.559	98.041	97.262
Central & South America	24.660	25.701	25.284	26.869	NA
Antarctica	0.003	0.004	0.003	0.001	NA
Antigua and Barbuda	0.009	0.010	0.009	0.009	NA
Argentina	3.273	3.352	3.302	3.349	NA

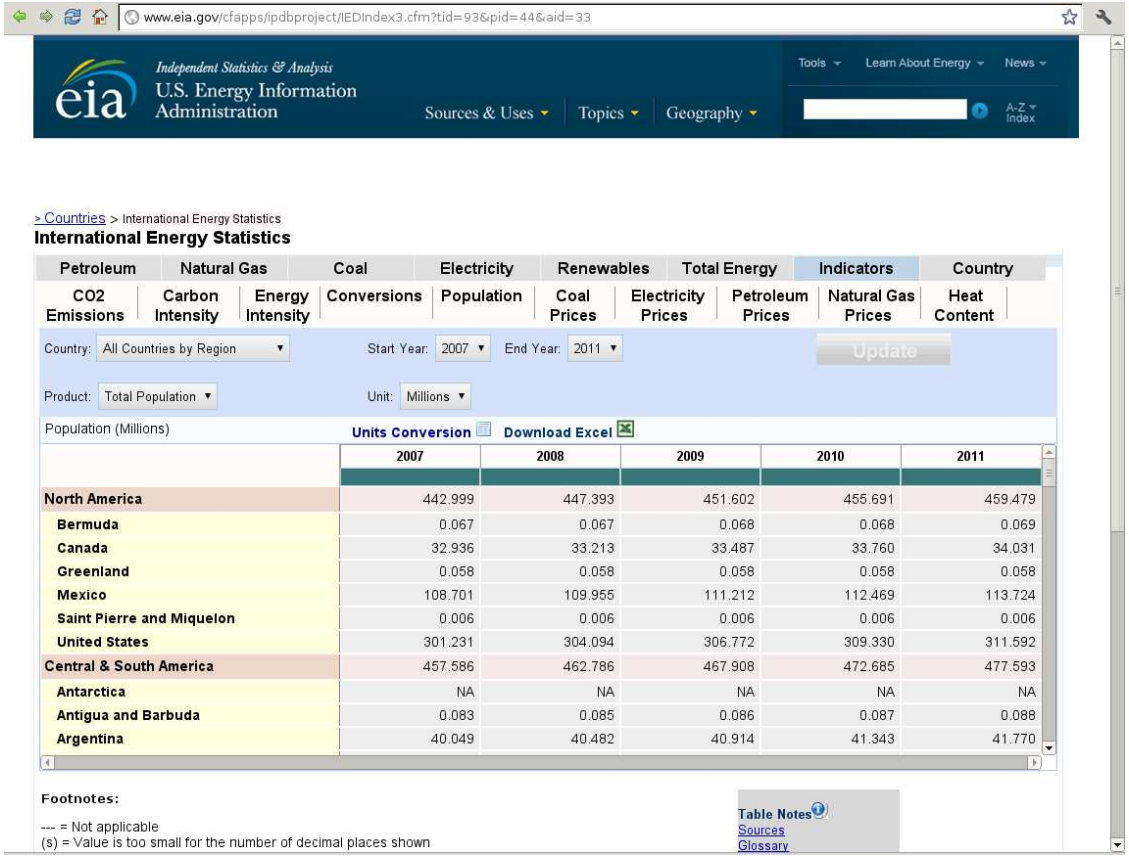
Note: It might have been interesting to set the “Start Year” to 1980

Obtain the data 2

Go to the URLs. Click Download Excel.

International Energy Statistics, Population

http://www.eia.gov/cfapps/ipdbproject/IEDIndex3.cfm?tid=93&pid=44&aid=33



	2007	2008	2009	2010	2011
North America	442.999	447.393	451.602	455.691	459.479
Bermuda	0.067	0.067	0.068	0.068	0.069
Canada	32.936	33.213	33.487	33.760	34.031
Greenland	0.058	0.058	0.058	0.058	0.058
Mexico	108.701	109.955	111.212	112.469	113.724
Saint Pierre and Miquelon	0.006	0.006	0.006	0.006	0.006
United States	301.231	304.094	306.772	309.330	311.592
Central & South America	457.586	462.786	467.908	472.685	477.593
Antarctica	NA	NA	NA	NA	NA
Antigua and Barbuda	0.083	0.085	0.086	0.087	0.088
Argentina	40.049	40.482	40.914	41.343	41.770

Note: It might have been interesting to set the “Start Year” to 1980

Clean the data 1

- In gnumeric save as CSV
- In vim, delete the first line with the table name
- Delete 2 lines that look like , , , , , ,
- Add the column name “Country” to the first line

```
"Table: Total Primary Energy Consumption (Quadrillion Btu)",,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
,,1980,1981,1982,1983,1984,1985,1986,1987,1988,1989,1990,1991,1992,1993,1994,1995,1996,1997,1
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
"North America",,91.60003,89.82128,86.64257,86.22497,90.60321,90.96346,91.24179,94.12101,98.3
```

Exploratory data analysis 1

```
$ R

R version 2.11.1 (2010-05-31)
Copyright (C) 2010 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

Exploratory data analysis 2

```
> EnergyConsumption <-
read.csv("Total_Primary_Energy_Consumption_(Quadrillion_Btu).csv",
na.strings="--",as.is=TRUE)
> Population <-
read.csv("Population_(Millions).csv",
na.strings="--",as.is=TRUE)
> Population <-
+ read.csv("Population_(Millions).csv",
+ na.strings="--",as.is=TRUE)
> class(EnergyConsumption)
[1] "data.frame"
> class(Population)
[1] "data.frame"
```

Exploratory data analysis 3

```
> head(EnergyConsumption)
                   Country  X     X2007     X2008     X2009     X2010    X2011
1            North America NA 122.86815 120.10351 114.62372 118.34544       NA
2                  Bermuda NA   0.01018   0.00981   0.00971   0.00955       NA
3                   Canada NA  14.02585  13.54625  13.09711  13.00072 13.49477
4                Greenland NA   0.00826   0.00917   0.00893   0.00893       NA
5                   Mexico NA   7.52670   7.26255   6.94764   7.28397   7.8079
6 Saint Pierre and Miquelon NA   0.00122   0.00123   0.00122   0.00122       NA
> tail(EnergyConsumption)
                 Country  X     X2007     X2008     X2009     X2010 X2011
227                Tonga NA   0.00252   0.00253   0.00252   0.00252    NA
228 U.S. Pacific Islands NA   0.00450   0.00507   0.00441   0.00513    NA
229              Vanuatu NA   0.00158   0.00143   0.00175   0.00175    NA
230              Vietnam NA   1.41820   1.63544   1.67657   1.86492    NA
231          Wake Island NA   0.01986   0.01888   0.01932   0.01916    NA
232                World NA 482.86461 490.68974 487.58009 510.55101    NA
> nrow(EnergyConsumption)
[1] 232
> ncol(EnergyConsumption)
[1] 7
```

Exploratory data analysis 4

```
> options(width = 90)
> head(Population)
                    Country  X     X2007     X2008     X2009     X2010     X2011
1             North America NA 442.99861 447.39306 451.60196 455.69066 459.47897
2                   Bermuda NA   0.06692   0.06739   0.06784   0.06827   0.06868
3                    Canada NA  32.93596  33.21270  33.48721  33.75974  34.03059
4                 Greenland NA   0.05753   0.05756   0.05760   0.05764   0.05767
5                    Mexico NA 108.70089 109.95540 111.21179 112.46886 113.72423
6 Saint Pierre and Miquelon NA   0.00610   0.00605   0.00600   0.00594   0.00589
> tail(Population)
                 Country  X      X2007      X2008      X2009      X2010      X2011
227                Tonga NA    0.10441    0.10488    0.10529    0.10563    0.10592
228 U.S. Pacific Islands NA         NA         NA         NA         NA         NA
229              Vanuatu NA    0.22882    0.23430    0.23978    0.24525    0.25072
230              Vietnam NA   86.51885   87.55836   88.57676   89.57113   90.54939
231          Wake Island NA         NA         NA         NA         NA         NA
232                World NA 6630.84486 6707.59554 6785.78950 6863.18955 6940.14288
> nrow(Population)
[1] 232
> ncol(Population)
[1] 7
```

Exploratory data analysis 5

Subsetting.

```
> Population[c(1,8:10),]
                   Country  X     X2007     X2008     X2009     X2010     X2011
1            North America NA 442.99861 447.39306 451.60196 455.69066 459.47897
8  Central & South America NA 457.58593 462.78618 467.90774 472.68456 477.59348
9               Antarctica NA        NA        NA        NA        NA        NA
10     Antigua and Barbuda NA   0.08343   0.08452   0.08563   0.08675   0.08788
>  Population[is.na(Population$X2011),]
                         Country  X X2007 X2008 X2009 X2010 X2011
9                     Antarctica NA    NA    NA    NA    NA    NA
66         Former Czechoslovakia NA    NA    NA    NA    NA    NA
67  Former Serbia and Montenegro NA    NA    NA    NA    NA    NA
68             Former Yugoslavia NA    NA    NA    NA    NA    NA
71                 Germany, East NA    NA    NA    NA    NA    NA
72                 Germany, West NA    NA    NA    NA    NA    NA
101              Former U.S.S.R. NA    NA    NA    NA    NA    NA
199          Hawaiian Trade Zone NA    NA    NA    NA    NA    NA
228         U.S. Pacific Islands NA    NA    NA    NA    NA    NA
231                  Wake Island NA    NA    NA    NA    NA    NA
> numpopcols <- ncol(Population)
> Population[1,3:numpopcols]
     X2007    X2008   X2009    X2010   X2011
1 442.9986 447.3931 451.602 455.6907 459.479
```

Exploratory data analysis 6

```
> EnergyConsumption[1,]
        Country  X    X2007    X2008    X2009    X2010 X2011
1 North America NA 122.8681 120.1035 114.6237 118.3454    NA
> sapply(EnergyConsumption[1,],class)
    Country           X       X2007       X2008       X2009       X2010       X2011
"character"   "logical"   "numeric"   "numeric"   "numeric"   "numeric" "character"
> Population[1,]
        Country  X     X2007     X2008     X2009     X2010     X2011
1 North America NA 442.99861 447.39306 451.60196 455.69066 459.47897
> sapply(Population[1,],class)
    Country           X       X2007       X2008       X2009       X2010       X2011
"character"   "logical" "character" "character" "character" "character" "character"
```

Clean the data 2

```
> Population[,3:numpopcols] <- lapply(Population[,3:numpopcols],as.numeric)
Warning messages:
1: In lapply(Population[, 3:numpopcols], as.numeric) :
  NAs introduced by coercion
2: In lapply(Population[, 3:numpopcols], as.numeric) :
  NAs introduced by coercion
3: In lapply(Population[, 3:numpopcols], as.numeric) :
  NAs introduced by coercion
4: In lapply(Population[, 3:numpopcols], as.numeric) :
  NAs introduced by coercion
5: In lapply(Population[, 3:numpopcols], as.numeric) :
  NAs introduced by coercion
> sapply(Population[1,],class)
    Country           X       X2007       X2008       X2009       X2010       X2011
"character"   "logical"   "numeric"   "numeric"   "numeric"   "numeric"   "numeric"
```

Exploratory data analysis 7

More subsetting.

```
> aggregates <- c("North America", "Central & South America", "Europe", "Eurasia",
"Middle East", "Africa", "Asia & Oceania", "World")
> head(EnergyConsumption[EnergyConsumption$Country %in% aggregates,])
                    Country  X     X2007     X2008     X2009     X2010 X2011
1             North America NA 122.86815 120.10351 114.62372 118.34544    NA
8   Central & South America NA  24.66038  25.70090  25.28353  26.86867    NA
54                   Europe NA  86.69852  85.64615  81.21917  83.82449    NA
96                  Eurasia NA  44.02493  45.06215  40.37198  42.83648    NA
113             Middle East NA  23.94363  26.09924  27.44746  28.73368    NA
128                  Africa NA  14.99456  16.09403  15.92452  16.32675    NA
> head(subset(EnergyConsumption,Country %in% aggregates))
                    Country  X     X2007     X2008     X2009     X2010 X2011
1             North America NA 122.86815 120.10351 114.62372 118.34544    NA
8   Central & South America NA  24.66038  25.70090  25.28353  26.86867    NA
54                   Europe NA  86.69852  85.64615  81.21917  83.82449    NA
96                  Eurasia NA  44.02493  45.06215  40.37198  42.83648    NA
113             Middle East NA  23.94363  26.09924  27.44746  28.73368    NA
128                  Africa NA  14.99456  16.09403  15.92452  16.32675    NA
> head(subset(EnergyConsumption,Country %in% aggregates)[,c(1,3:7)])
                    Country     X2007     X2008     X2009     X2010 X2011
1             North America 122.86815 120.10351 114.62372 118.34544    NA
8   Central & South America  24.66038  25.70090  25.28353  26.86867    NA
54                   Europe  86.69852  85.64615  81.21917  83.82449    NA
96                  Eurasia  44.02493  45.06215  40.37198  42.83648    NA
113             Middle East  23.94363  26.09924  27.44746  28.73368    NA
128                  Africa  14.99456  16.09403  15.92452  16.32675    NA
```

Define the Question

Source: Global Sustainable Energy: Past, Present and Future
`https://www.coursera.org/course/globalenergy`
What is the geographical location of the countries that consume less than 219 kWh per year? Note that these are the countries that are essentially living on the equivalent of human power as their source of energy.
Also, what is the geographical location of the countries that consume less than 1,314 kWh per year? Note that these are the countries that are essentially living on the equivalent of oxen power as their source of energy.
How many countries are there in this list?
In terms of population, what is the largest country in this list?

Define the Question

Source: Global Sustainable Energy: Past, Present and Future
`https://www.coursera.org/course/globalenergy`
What is the geographical location of the countries that consume less than 219 kWh per year? Note that these are the countries that are essentially living on the equivalent of human power as their source of energy.
Also, what is the geographical location of the countries that consume less than 1,314 kWh per year? Note that these are the countries that are essentially living on the equivalent of oxen power as their source of energy.
How many countries are there in this list?
In terms of population, what is the largest country in this list?

Define the Question

Source: Global Sustainable Energy: Past, Present and Future
`https://www.coursera.org/course/globalenergy`
What is the geographical location of the countries that consume less than 219 kWh per year? Note that these are the countries that are essentially living on the equivalent of human power as their source of energy.
Also, what is the geographical location of the countries that consume less than 1,314 kWh per year? Note that these are the countries that are essentially living on the equivalent of oxen power as their source of energy.
How many countries are there in this list?
In terms of population, what is the largest country in this list?

Clean the data 3

```
> EConPC <- merge(EnergyConsumption[,c(1,3:7)],Population[,c(1,3:7)],
+ by.x="Country",by.y="Country",sort = FALSE,suffixes=c("ECon","Pop"))
> head(EConPC)
                   Country X2007ECon X2008ECon X2009ECon X2010ECon X2011ECon  X2007Pop
1            North America 122.86815 120.10351 114.62372 118.34544        NA 442.99861
2                  Bermuda   0.01018   0.00981   0.00971   0.00955        NA   0.06692
3                   Canada  14.02585  13.54625  13.09711  13.00072  13.49477  32.93596
4                Greenland   0.00826   0.00917   0.00893   0.00893        NA   0.05753
5                   Mexico   7.52670   7.26255   6.94764   7.28397    7.8079 108.70089
6 Saint Pierre and Miquelon   0.00122   0.00123   0.00122   0.00122        NA   0.00610
   X2008Pop  X2009Pop  X2010Pop  X2011Pop
1 447.39306 451.60196 455.69066 459.47897
2   0.06739   0.06784   0.06827   0.06868
3  33.21270  33.48721  33.75974  34.03059
4   0.05756   0.05760   0.05764   0.05767
5 109.95540 111.21179 112.46886 113.72423
6   0.00605   0.00600   0.00594   0.00589
```

Clean the data 4

```
> CurrContinent <- aggregates[1]
> for(i in 1:length(EConPC$Country)) {
+   if (EConPC$Country[i] %in% aggregates) {
+     EConPC$Continent[i] <- EConPC$Country[i]
+     CurrContinent <- EConPC$Country[i]
+   } else {
+     EConPC$Continent[i] <- CurrContinent
+   }
+ }
> head(EConPC)
                    Country X2007ECon X2008ECon X2009ECon X2010ECon X2011ECon  X2007Pop
1             North America 122.86815 120.10351 114.62372 118.34544        NA 442.99861
2                   Bermuda   0.01018   0.00981   0.00971   0.00955        NA   0.06692
3                    Canada  14.02585  13.54625  13.09711  13.00072  13.49477  32.93596
4                 Greenland   0.00826   0.00917   0.00893   0.00893        NA   0.05753
5                    Mexico   7.52670   7.26255   6.94764   7.28397    7.8079 108.70089
6 Saint Pierre and Miquelon   0.00122   0.00123   0.00122   0.00122        NA   0.00610
   X2008Pop  X2009Pop  X2010Pop  X2011Pop     Continent
1 447.39306 451.60196 455.69066 459.47897 North America
2   0.06739   0.06784   0.06827   0.06868 North America
3  33.21270  33.48721  33.75974  34.03059 North America
4   0.05756   0.05760   0.05764   0.05767 North America
5 109.95540 111.21179 112.46886 113.72423 North America
6   0.00605   0.00600   0.00594   0.00589 North America
```

Clean the data 5

Compute Energy per capita.
1 quad = 1 Quadrillion BTU = 293,083,000,000 kWh

http://en.wikipedia.org/wiki/Quad_(unit)

EIA's energy converter is helpful (But gives 293,071,111,111 kWh instead):

http://www.iea.org/stats/unit.asp

```
> EConPC$EConPC2007 <- (EConPC[,2] * 293083000000) / (EConPC[,2+5] * 1000000)
> EConPC$EConPC2008 <- (EConPC[,3] * 293083000000) / (EConPC[,3+5] * 1000000)
> EConPC$EConPC2009 <- (EConPC[,4] * 293083000000) / (EConPC[,4+5] * 1000000)
> EConPC$EConPC2010 <- (EConPC[,5] * 293083000000) / (EConPC[,5+5] * 1000000)
> head(EConPC,5)
        Country X2007ECon X2008ECon X2009ECon X2010ECon X2011ECon  X2007Pop  X2008Pop
1 North America 122.86815 120.10351 114.62372 118.34544        NA 442.99861 447.39306
2       Bermuda   0.01018   0.00981   0.00971   0.00955        NA   0.06692   0.06739
3        Canada  14.02585  13.54625  13.09711  13.00072  13.49477  32.93596  33.21270
4     Greenland   0.00826   0.00917   0.00893   0.00893        NA   0.05753   0.05756
5        Mexico   7.52670   7.26255   6.94764   7.28397    7.8079 108.70089 109.95540
   X2009Pop  X2010Pop  X2011Pop     Continent EConPC2007 EConPC2008 EConPC2009 EConPC2010
1 451.60196 455.69066 459.47897 North America   81288.21   78678.68   74389.10   76115.31
2   0.06784   0.06827   0.06868 North America   44584.35   42664.26   41949.23   40998.13
3  33.48721  33.75974  34.03059 North America  124810.03  119537.88  114627.06  112864.91
4   0.05760   0.05764   0.05767 North America   42080.06   46691.65   45438.04   45406.51
5 111.21179 112.46886 113.72423 North America   20293.74   19358.12   18309.53   18981.32
```

Exploratory data analysis 8

```
> library(plyr)
> below219 <- subset(EConPC,EConPC2010 <= 219 & !is.na(Country))[c(1,5,10,16,12)]
> below1314 <- subset(EConPC,EConPC2010 <= 1314 & !is.na(Country))[c(1,5,10,16,12)]
> arrange(below219, desc(EConPC2010))
  Country X2010ECon X2010Pop EConPC2010 Continent
1 Burundi   0.00468  9.86312   139.0664    Africa
2    Chad   0.00364 10.54346   101.1833    Africa
> nrow(below219)
[1] 2
> nrow(below1314)
[1] 27
```

Exploratory data analysis 9

```
> arrange(below1314, desc(EConPC2010))
                   Country X2010ECon X2010Pop EConPC2010              Continent
1                 Cambodia   0.06118 14.45368  1240.5711         Asia & Oceania
2              Sierra Leone   0.02034  5.24570  1136.4181                 Africa
3               Gambia, The   0.00677  1.75546  1130.2860                 Africa
4               Afghanistan   0.10445 29.12073  1051.2277         Asia & Oceania
5             Guinea-Bissau   0.00557  1.56513  1043.0267                 Africa
6                      Togo   0.02309  6.58724  1027.3326                 Africa
7                     Haiti   0.03097  9.64892   940.7043 Central & South America
8                     Nepal   0.08436 28.95185   853.9863         Asia & Oceania
9                  Tanzania   0.12017 44.28820   795.2408                 Africa
10                  Comoros   0.00186  0.70612   772.0138                 Africa
11 Timor-Leste (East Timor)   0.00262  1.08767   705.9839         Asia & Oceania
12                   Guinea   0.02447 10.32403   694.6649                 Africa
13                  Liberia   0.00753  3.68508   598.8784                 Africa
14                   Uganda   0.06250 31.50723   581.3804                 Africa
15                   Malawi   0.03061 15.44750   580.7587                 Africa
16         Congo (Kinshasa)   0.11299 69.85129   474.0850                 Africa
17             Burkina Faso   0.02599 16.24181   468.9888                 Africa
18                 Ethiopia   0.13678 86.04293   465.9057                 Africa
19               Madagascar   0.03292 20.84662   462.8229                 Africa
20 Central African Republic   0.00609  4.84493   368.4007                 Africa
21                  Eritrea   0.00701  5.79298   354.6554                 Africa
22                    Niger   0.01834 15.27002   352.0062                 Africa
23                  Somalia   0.01125  9.76789   337.5533                 Africa
24                   Rwanda   0.01219 11.05598   323.1447                 Africa
25                     Mali   0.01256 14.58261   252.4323                 Africa
26                  Burundi   0.00468  9.86312   139.0664                 Africa
27                     Chad   0.00364 10.54346   101.1833                 Africa
```

Exploratory data analysis 10

```
> arrange(below1314, desc(X2010Pop))
                   Country X2010ECon X2010Pop EConPC2010                Continent
1                 Ethiopia   0.13678 86.04293   465.9057                   Africa
2          Congo (Kinshasa)   0.11299 69.85129   474.0850                   Africa
3                 Tanzania   0.12017 44.28820   795.2408                   Africa
4                   Uganda   0.06250 31.50723   581.3804                   Africa
5              Afghanistan   0.10445 29.12073  1051.2277           Asia & Oceania
6                    Nepal   0.08436 28.95185   853.9863           Asia & Oceania
7               Madagascar   0.03292 20.84662   462.8229                   Africa
8             Burkina Faso   0.02599 16.24181   468.9888                   Africa
9                   Malawi   0.03061 15.44750   580.7587                   Africa
10                   Niger   0.01834 15.27002   352.0062                   Africa
11                    Mali   0.01256 14.58261   252.4323                   Africa
12                Cambodia   0.06118 14.45368  1240.5711           Asia & Oceania
13                  Rwanda   0.01219 11.05598   323.1447                   Africa
14                    Chad   0.00364 10.54346   101.1833                   Africa
15                  Guinea   0.02447 10.32403   694.6649                   Africa
16                  Burundi   0.00468  9.86312   139.0664                   Africa
17                  Somalia   0.01125  9.76789   337.5533                   Africa
18                   Haiti   0.03097  9.64892   940.7043 Central & South America
19                    Togo   0.02309  6.58724  1027.3326                   Africa
20                  Eritrea   0.00701  5.79298   354.6554                   Africa
21             Sierra Leone   0.02034  5.24570  1136.4181                   Africa
22 Central African Republic   0.00609  4.84493   368.4007                   Africa
23                  Liberia   0.00753  3.68508   598.8784                   Africa
24              Gambia, The   0.00677  1.75546  1130.2860                   Africa
25            Guinea-Bissau   0.00557  1.56513  1043.0267                   Africa
26 Timor-Leste (East Timor)   0.00262  1.08767   705.9839           Asia & Oceania
27                  Comoros   0.00186  0.70612   772.0138                   Africa
```

Exploratory data analysis 11

```
> arrange(below1314, desc(X2010Pop))[,1]
 [1] "Ethiopia"                 "Congo (Kinshasa)"         "Tanzania"
 [4] "Uganda"                   "Afghanistan"              "Nepal"
 [7] "Madagascar"               "Burkina Faso"             "Malawi"
[10] "Niger"                    "Mali"                     "Cambodia"
[13] "Rwanda"                   "Chad"                     "Guinea"
[16] "Burundi"                  "Somalia"                  "Haiti"
[19] "Togo"                     "Eritrea"                  "Sierra Leone"
[22] "Central African Republic" "Liberia"                  "Gambia, The"
[25] "Guinea-Bissau"            "Timor-Leste (East Timor)" "Comoros"
> sort(below1314[,1])
 [1] "Afghanistan"              "Burkina Faso"             "Burundi"
 [4] "Cambodia"                 "Central African Republic" "Chad"
 [7] "Comoros"                  "Congo (Kinshasa)"         "Eritrea"
[10] "Ethiopia"                 "Gambia, The"              "Guinea"
[13] "Guinea-Bissau"            "Haiti"                    "Liberia"
[16] "Madagascar"               "Malawi"                   "Mali"
[19] "Nepal"                    "Niger"                    "Rwanda"
[22] "Sierra Leone"             "Somalia"                  "Tanzania"
[25] "Timor-Leste (East Timor)" "Togo"                     "Uganda"
```

Clean the data 6

```
> EConPC$HSPC2010 <- EConPC$EConPC2010 / 219
> EConPC$OSPC2010 <- EConPC$EConPC2010 / 1314
> mostslaves <- subset(EConPC,! Country %in% aggregates)[c(1,16,17,18,12)]
> head(arrange(mostslaves, desc(EConPC2010)),20)
                Country EConPC2010  HSPC2010  OSPC2010               Continent
1             Gibraltar  646243.96 2950.8856 491.81427                  Europe
2  Virgin Islands,  U.S.  629189.38 2873.0109 478.83515 Central & South America
3    Trinidad and Tobago  223324.05 1019.7445 169.95742 Central & South America
4   Netherlands Antilles  210639.39  961.8237 160.30395 Central & South America
5   United Arab Emirates  208292.21  951.1060 158.51767             Middle East
6               Iceland  194895.31  889.9329 148.32215                  Europe
7             Singapore  192264.61  877.9206 146.32010          Asia & Oceania
8                 Qatar  188271.83  859.6887 143.28145             Middle East
9                Kuwait  147911.17  675.3934 112.56557             Middle East
10              Bahrain  138482.26  632.3391 105.38985             Middle East
11               Norway  117097.57  534.6921  89.11535                  Europe
12           Luxembourg  117070.62  534.5690  89.09484                  Europe
13               Canada  112864.91  515.3649  85.89415           North America
14               Brunei  109649.23  500.6814  83.44690          Asia & Oceania
15        United States   92891.55  424.1624  70.69373           North America
16         Saudi Arabia   89398.82  408.2138  68.03564             Middle East
17                 Oman   85623.27  390.9738  65.16231             Middle East
18            Australia   79162.48  361.4725  60.24542          Asia & Oceania
19              Belgium   77269.00  352.8265  58.80441                  Europe
20          Netherlands   75441.22  344.4805  57.41341                  Europe
```

Statistical prediction/modeling

No statistical predication or modeling was performed for this energy slaves data analysis.

Interpret results

- Chad and Burundi, two African nations, are the only countries in the world where per capita energy consumption is less than the energy equivalent of the power of a human being, a so-called “energy slave”
- There are 27 nations around the world whose per capita energy consumption is less than the energy equivalent of the power of an ox. The largest of these in terms of population is Ethiopia with 86 million people.
- The remaining 197 nations of the world have at least an ox worth of energy consumption available to them. Gibraltar has the most energy available to them with the equivalent of 2950 human-scale energy slaves or 491 ox-scale energy slaves.

Exploratory data analysis 12

```
> nrow(mostslaves)
[1] 224
> 224-27
[1] 197
```

Challenge results

- The data on population and energy are impossible to be 100% accurate and precise as they are moving targets and there is no way to ensure that centrally collected data is accurate. Moreover, central data collection systems are subject to various biases and omissions of unknown dimension and character.
- There are other data sources with comparable information (World Bank, UN, etc.). We ought to repeat the analysis on those data to assess if the range of accuracy from the different sources is comparable. If discrepancies exist between the different data sets an assessment to determine which sources are most accurate could affect interpretation of the data.

Synthesize/write up results

- Energy consumption varies widely around the world. Gibraltar has the equivalent of 2950 human-scale “energy slaves” while nations such as Chad and Burundi have less than one. There are 27 nations in the world who do not have an oxen-scale energy slave at their disposal. The largest country with less than an oxen worth of energy consumption is Ethiopia with 86 million people.

Create reproducible code

See `http://www.CJFearnley.com/Data.Analysis.PLUG.July2013.R`

Conclusion

- The two coursera courses *Computing for Data Analysis* with Roger Peng (`http://www.coursera.org/course/compdata`) and *Data Analysis* with Jeff Leek (`http://www.coursera.org/course/dataanalysis`) provide a great introduction to data analysis using the R Programming language.
- R is included with all major Linux distributions
- So Enjoy!

Conclusion

- The two coursera courses *Computing for Data Analysis* with Roger Peng (`http://www.coursera.org/course/compdata`) and *Data Analysis* with Jeff Leek (`http://www.coursera.org/course/dataanalysis`) provide a great introduction to data analysis using the R Programming language.
- R is included with all major Linux distributions
- So Enjoy!

Conclusion

- The two coursera courses *Computing for Data Analysis* with Roger Peng (`http://www.coursera.org/course/compdata`) and *Data Analysis* with Jeff Leek (`http://www.coursera.org/course/dataanalysis`) provide a great introduction to data analysis using the R Programming language.
- R is included with all major Linux distributions
- So Enjoy!

Thank You

Thank You!

Any Questions?

On-line version of this presentation:
`http://www.CJFearnley.com/Data.Analysis.PLUG.July2013.pdf`