Three types of dataset

* **Cross-section data**: data on one or more variables collected ***at the same point in time*** (i.e., **multiple subjects** or individuals at the **same time**).

**Structure**

A: x= Profit (year 2015, billions $), i = {Apple, Microsoft, GE, IBM, etc. }

xi : x1 (AAPL): 40

x2 (MSFT): 50

x3 (GE): 75

x4 (IBM): 100

------

B: x = GDP (year 2016, billions $) , I = { US, China, Japan, Germany, etc}

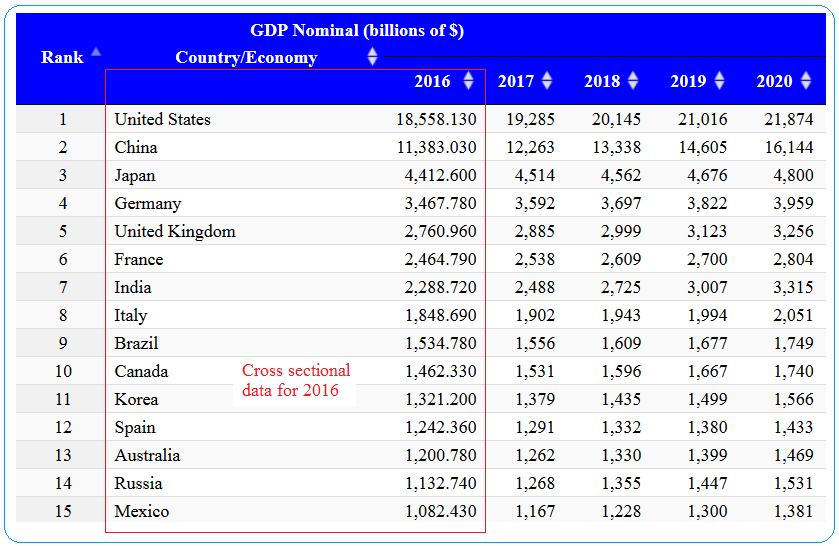
xi : x1(US): 18,558.130

x2 (China): 11,383.030

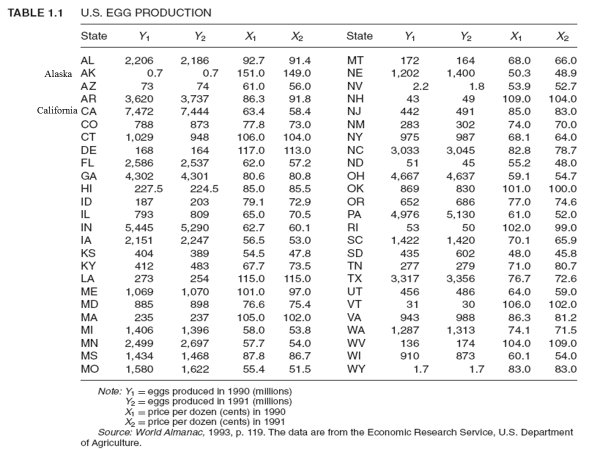
x3 (Japan): 4,412.600

x4 (Germany): 3,467.780

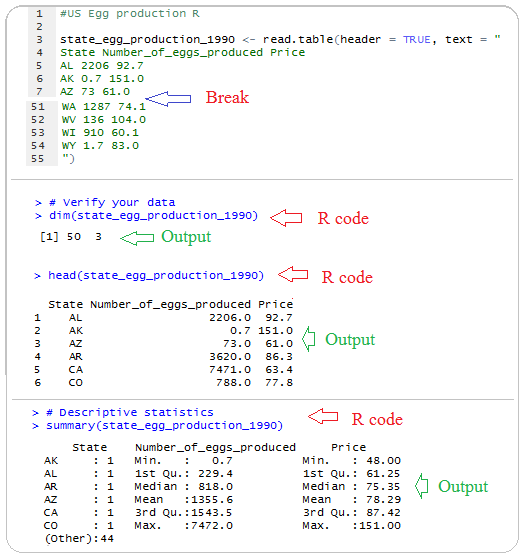
-------

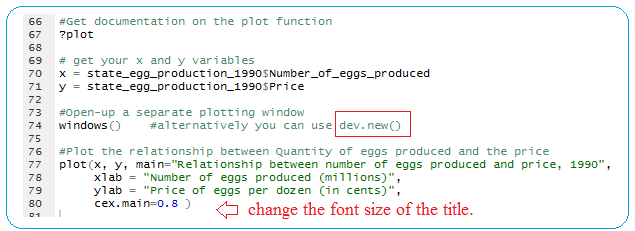
C:

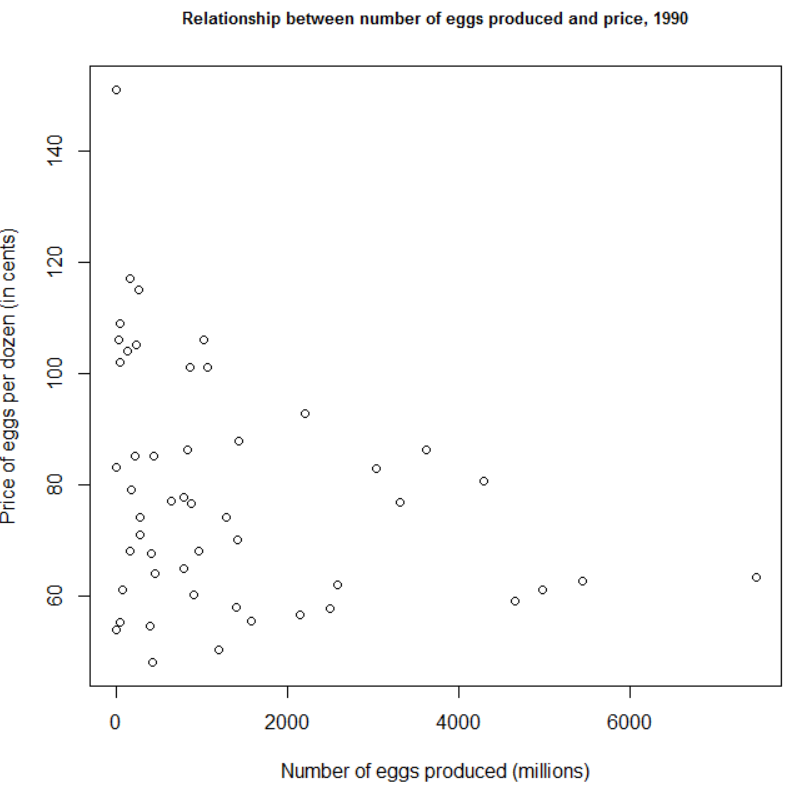
Source: <http://statisticstimes.com/economy/countries-by-projected-gdp.php>

D:

* For **each year** the data on the 50 states are cross-sectional data.

**R-Code:**



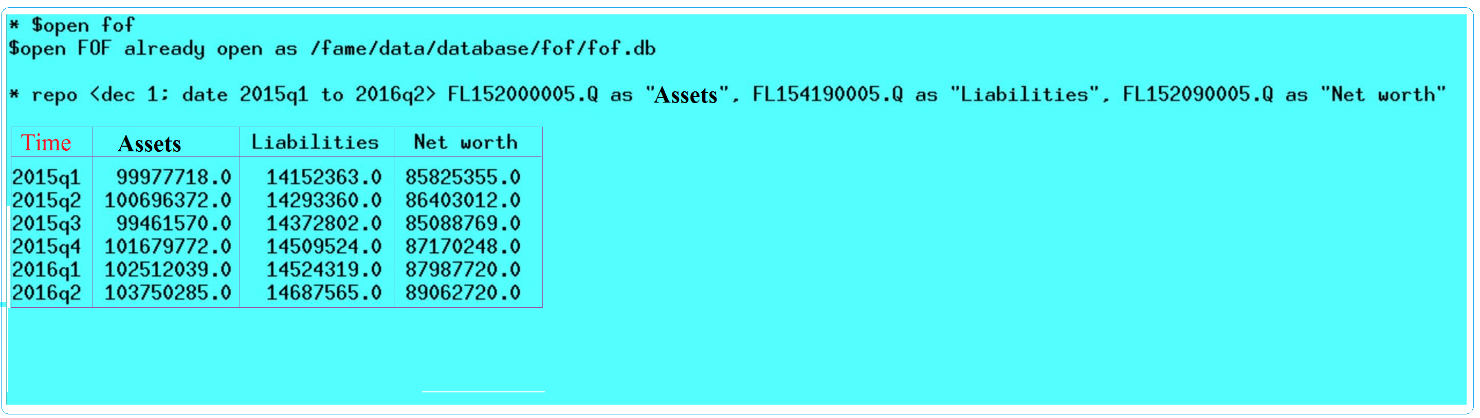


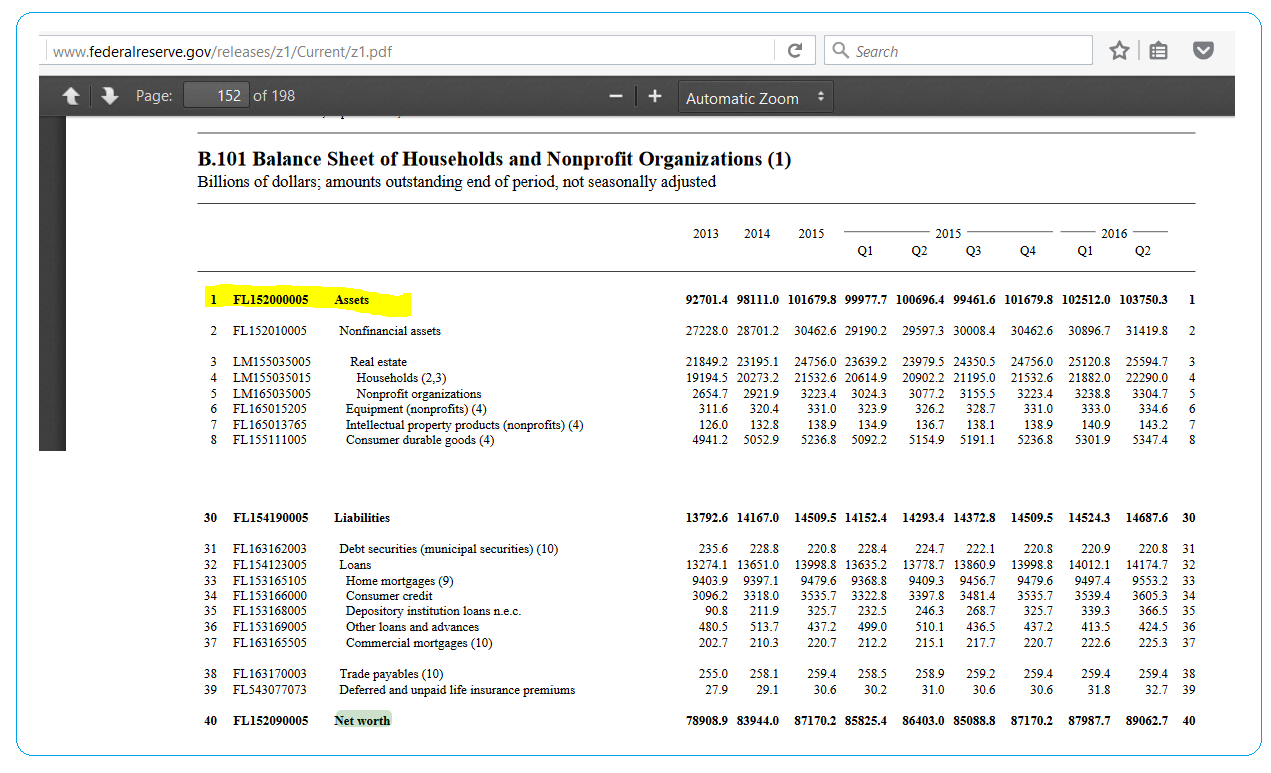
* **Time series data**: A time series is a set of observations on the values that a variable takes at different times. It is collected at **regular time intervals**, such as daily, weekly, monthly quarterly, etc.

**Structure and example**

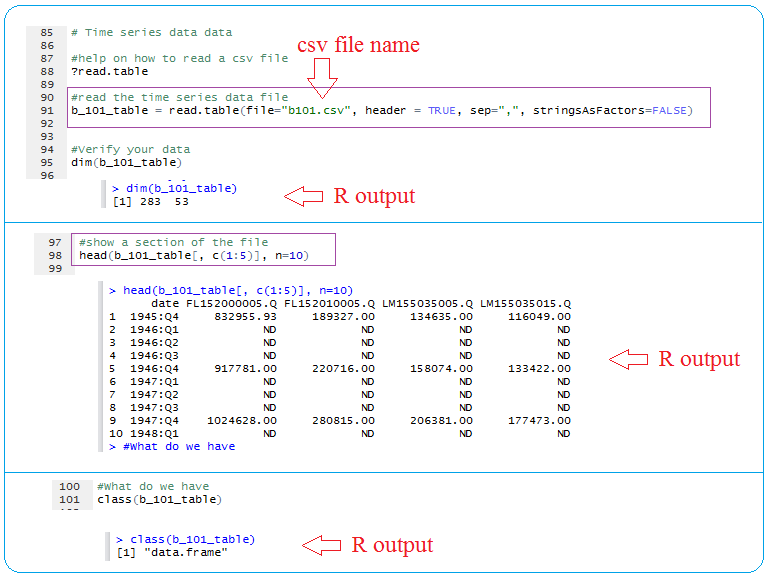
Example 1: Sales time series

****

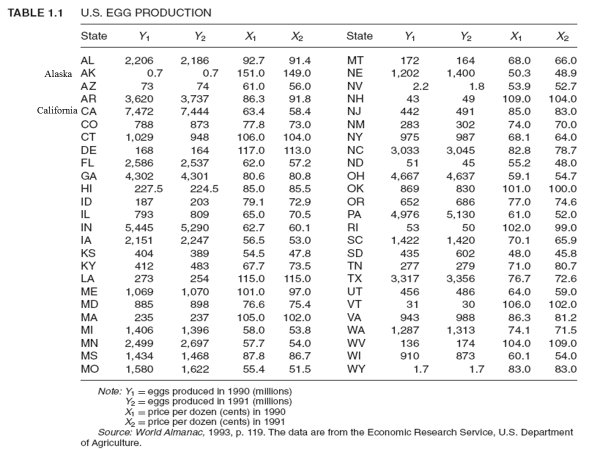
Example 2: Financial Accounts time series



**RCode**



**Pooled data**: In pooled, or combined, data are elements of ***both time series and cross-section*** *dat*a. The data in Table 1.1 are an example of pooled data. For each year we have 50 cross-sectional observations and for each state we have two-time series observations on prices and output of eggs, a total of 100 *pooled* (or combined) observations.



* **Panel, Longitudinal, or Micropanel Data**: This is a ***special type***of pooled data in which the ***same cross-sectional unit*** *(say, a family or a firm)* is surveyed over time.

Another definition: In pooled, or combined, data are elements of ***both time series and cross-section*** *dat*a. panel (longitudinal)

* + multiple subjects (individuals)
  + at different times, you have the same subject at different times, and you have many subjects at the same time; think of it as a table where rows are time points, and columns are subjects.

For example, part of a longitudinal dataset could contain specific students and their standardized test scores in six successive years.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Student Name** | **Grade 1 (2001) Raw Score** | **Grade 2 (2002) Raw Score** | **Grade 3 (2003) Raw Score** | **Grade 4 (2004)  Raw Score** | **Grade 5 (2005) Raw Score** | **Grade 6 (2006) Raw Score** |
| Mike | 339 | 350 | 361 | 366 | 381 | 390 |
| Jasmine | 332 | 343 | 350 | 351 | 351 | 355 |
| Thomas | 360 | 380 | 400 | 420 | 430 | 438 |

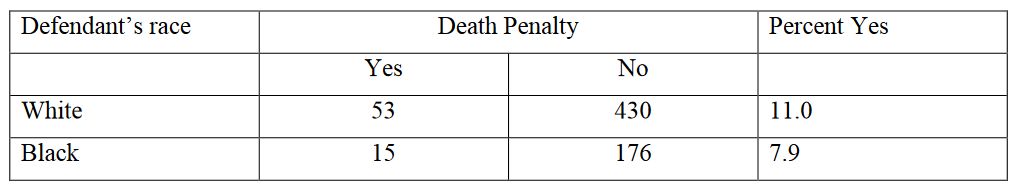
**The primary advantage of longitudinal databases is that they can measure change.** So we can estimate, for example, the effect of various factors on improvement in student achievement. We can also estimate the overall effectiveness of individual teachers by examining the performance of successive classes of students they teach, as well as examine the extent to which teacher effectiveness changes with experience or the composition of their class.

Source: <http://www.caldercenter.org/what-are-longitudinal-data>

**Motivations for Multilevel Models**

Consider the following the data from Agresti (1996). Researchers were interested in possible bias

in death penalty cases based on the defendant’s race. Here is a simple table examining defendant’s race and whether they were given the death penalty:

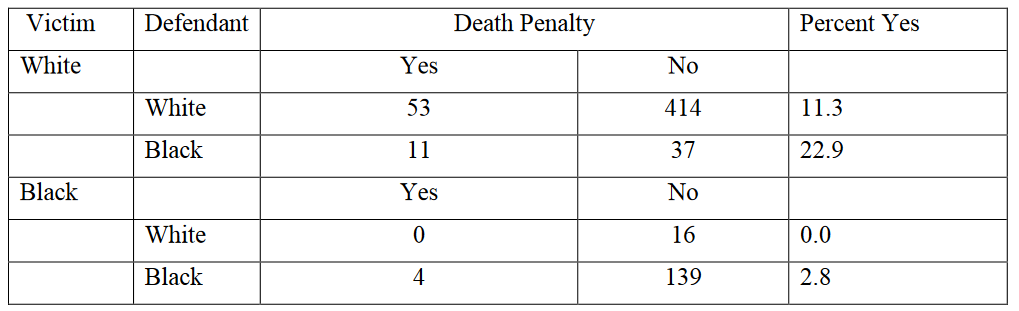


Based on these data, we would conclude – if anything – that there is a slight bias against White

defendants. However, there was also data on the victim’s race as well.

**Another view:**

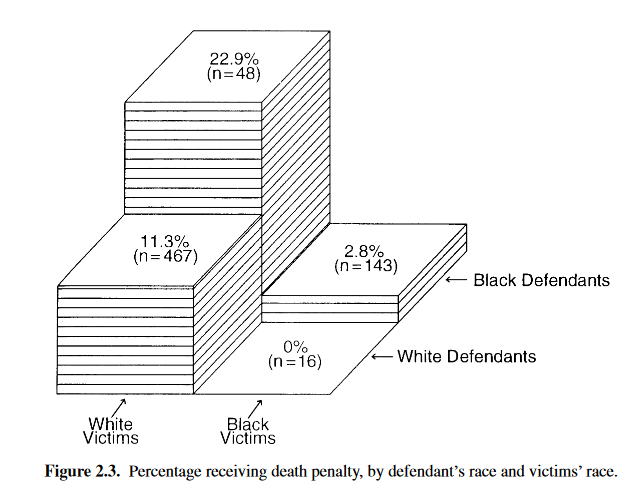
Let’s look at what the data look like when we disaggregate the data by victim’s race:



Now things look quite a bit different.

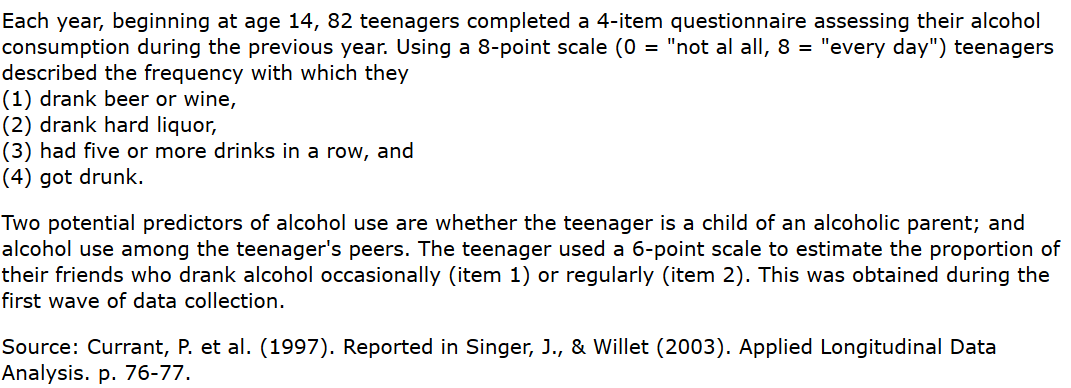
Once we take into account the victim’s race, a greater percentage of Black defendant’s are given the death penalty – regardless of victim’s race!

Source: <https://mregresion.files.wordpress.com/2012/08/agresti-introduction-to-categorical-data.pdf>



**Working with Panel dataset**

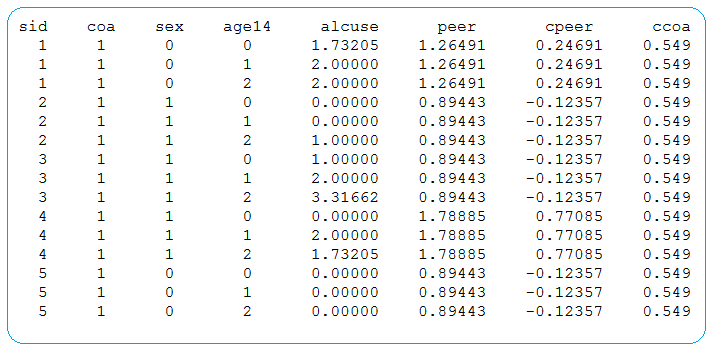
**Example1:**

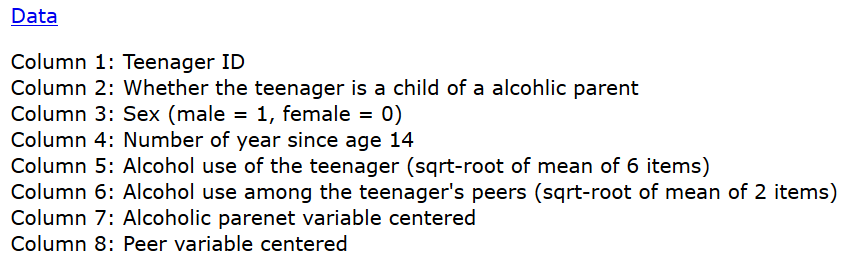


--

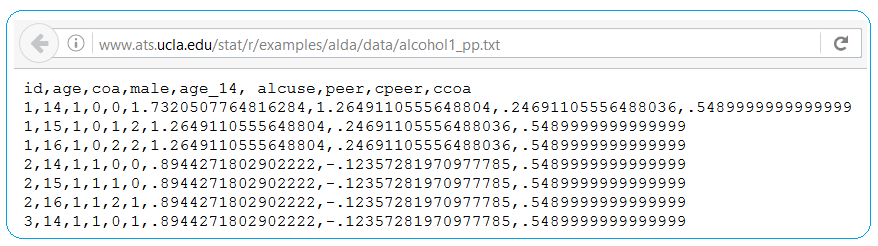
**The dataset**

<http://www.ats.ucla.edu/stat/r/examples/alda/data/alcohol1_pp.txt>



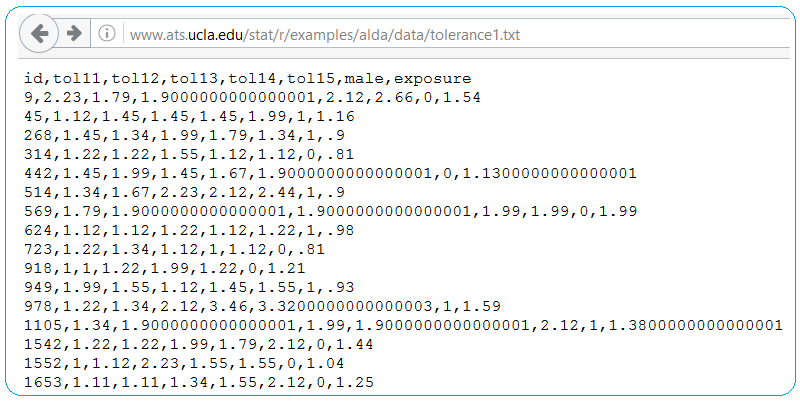


**Another presentation of the dataset**

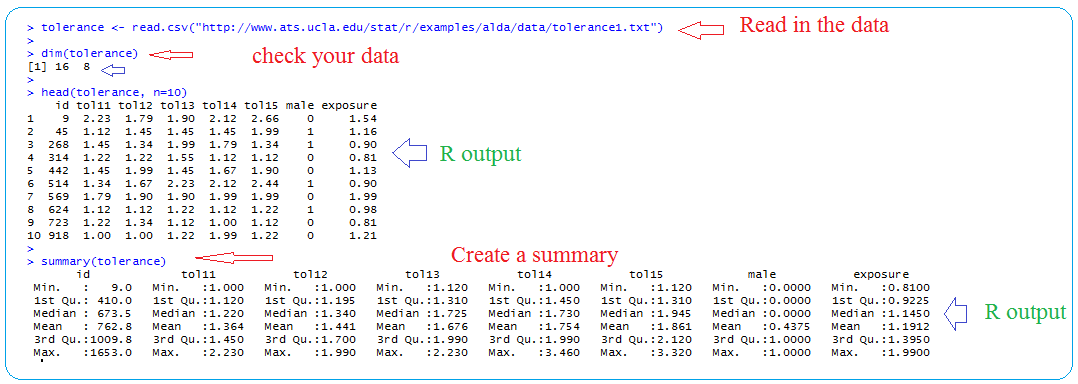


**Another example with R code**

<http://www.ats.ucla.edu/stat/r/examples/alda/data/tolerance1.txt>



**R code to read data**

****

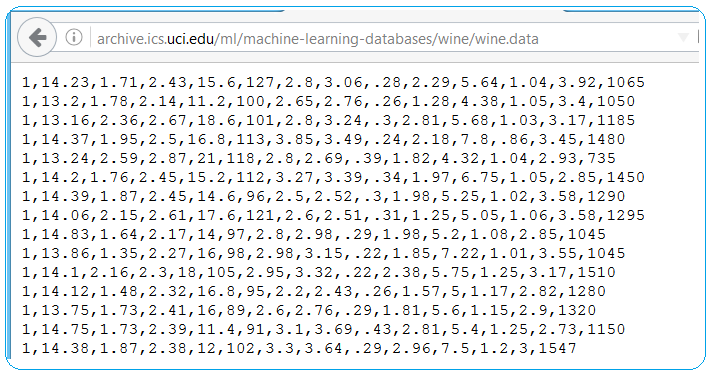
**See worked example on the Wine dataset from UCI data repository:**

[**https://github.com/wampeh1/ECOG\_314/blob/master/project1/lecture3\_project\_guide\_multivariate\_data\_analysis\_example.rmd**](https://github.com/wampeh1/ECOG_314/blob/master/project1/lecture3_project_guide_multivariate_data_analysis_example.rmd)

[**https://github.com/wampeh1/ECOG\_314/blob/master/project1/pdf/lecture3\_project\_guide\_multivariate\_data\_analysis\_example.pdf**](https://github.com/wampeh1/ECOG_314/blob/master/project1/pdf/lecture3_project_guide_multivariate_data_analysis_example.pdf)

**Data file:**

[**http://archive.ics.uci.edu/ml/datasets/Wine**](http://archive.ics.uci.edu/ml/datasets/Wine)

****

