Data Exploration:

Once you bring your data in STATA the next step is getting comfortable with your data.

Example:

We will begin by loading `hs0.dta`, a dataset saved in Stata's format. Stata data files end with the `dta` extension. Stata data files are loaded into memory using the `use` command.

```
. use http://www.ats.ucla.edu/stat/data/hs0, clear
```

Before we start our statistical exploration we will look at the data using the `describe, list` and `codebook` commands. Note that the variable `prgtype` is a string variable.

GETTING COMFORTABE WITH DATA

```
  ___  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   11.1   Co
> pyright 2009 StataCorp LP
  Statistics/Data Analysis            St
> ataCorp
                                      49
> 05 Lakeway Drive
                                      Co
> llege Station, Texas 77845 USA
                                      80
> 0-STATA-PC        http://www.stata.com
                                      97
> 9-696-4600        stata@stata.com
                                      97
> 9-696-4601 (fax)

Single-user Stata perpetual license:
       Serial number:  30110536678
         Licensed to:  Michelle Edwards
                       DRC ug

Notes:
      1.  (/m# option or -set memory-) 1
> 0.00 MB allocated to data
      2.  New update available; type -up
> date all-

. use http://www.ats.ucla.edu/stat/data/hs0, clear

. describe

Contains data from http://www.ats.ucla.edu/stat/data/hs0.dta
  obs:           200
 vars:            11                          12 Dec 2008 14:38
 size:        10,400 (99.9% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
gender          float  %9.0g
id              float  %9.0g
race            float  %12.0g      rl
ses             float  %9.0g       sl
schtyp          float  %9.0g
prgtype         str8   %9s
read            float  %9.0g                 reading score
write           float  %9.0g                 writing score
math            float  %9.0g                 math score
science         float  %9.0g                 science score
socst           float  %9.0g                 social studies score
-------------------------------------------------------------------------------
```

By typing `describe` in the command window we get the information about the variables and size of the file.

```
. codebook

gender                                                                  (unl
-------------------------------------------------------------------------------

                type:  numeric (float)

               range:  [1,2]                        units:  1
        unique values:  2                       missing .:  0/200

         tabulation:  Freq.  Value
                       91  1
                      109  2

-------------------------------------------------------------------------------
id                                                                      (unl
-------------------------------------------------------------------------------

                type:  numeric (float)

               range:  [1,200]                      units:  1
        unique values:  200                     missing .:  0/200

                mean:     100.5
            std. dev:   57.8792

          percentiles:        10%       25%       50%       75%       90%
                             20.5      50.5     100.5     150.5     180.5

-------------------------------------------------------------------------------
race                                                                    (unl
-------------------------------------------------------------------------------

                type:  numeric (float)
               label:  rl, but 1 nonmissing value is not labeled

               range:  [1,5]                        units:  1
        unique values:  5                       missing .:  0/200

         tabulation:  Freq.   Numeric  Label
                       24         1  hispanic
                       11         2  asian
                       20         3  african-amer
                      143         4  white
                        2         5
```

By typing codebook in the command window we get the information about each variable such as type of the variable, range of the variable, value labels, if there are any missing values.

Next, we will open a log file which will save all of the commands and the output in a text file. We can open a log file by typing:

```
. log using example1.txt, text replace
```

```
      name:  <unnamed>
       log:  C:\Documents and Settings\aurangze\My Documents\example1.txt
  log type:  text
 opened on:   8 Feb 2011, 14:53:27
```

STATA saved a new log file name `example1.txt` in the default directory, you can change the directory by using `cd` command. For example:

```
. cd "C:\Users\Desktop"
```

The basic descriptive statistics command in Stata is `summarize`. Along with `summarize`, we also show the `tabstat` and table commands for displaying descriptive statistics within groups. Type:

```
. summarize read math science write
```

in the command window we get the following descriptive stats:

```
. summarize read math science write

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        read |       200       52.23    10.25294         28         76
        math |       200      52.645    9.368448         33         75
     science |       195    51.66154    9.866026         26         74
       write |       200      52.775    9.478586         31         67

.
```

We can also use 'tabstat' command to get the summary statistics. This command is useful if we want to get the summary by grouping the cases.

```
. tabstat read write math, by(prgtype) stat(n mean sd)

Summary statistics: N, mean, sd
  by categories of: prgtype

 prgtype |      read     write      math
---------+------------------------------
academic |       105       105       105
         |   56.1619  56.25714  56.73333
         |  9.588779  7.943343  8.730216
---------+------------------------------
 general |        45        45        45
         |  49.75556  51.33333  50.02222
         |  9.234706  9.397775  7.442168
---------+------------------------------
  vocati |        50        50        50
         |      46.2     46.76     46.42
         |   8.90769  9.318754   7.95418
---------+------------------------------
   Total |       200       200       200
         |     52.23     52.775     52.645
         |  10.25294  9.478586  9.368448
-------------------------------------------
```

The tabulate command can produce one-way or two-way frequency tables. The `tab` command is a convenience command to produce multiple one-way frequency tables.

```
. tabstat write, by(prgtype) stat(n mean sd p25 p50 p75)

Summary for variables: write
     by categories of: prgtype

 prgtype |         N      mean        sd       p25       p50       p75
---------+--------------------------------------------------------------
academic |       105  56.25714  7.943343        52        59        62
 general |        45  51.33333  9.397775        44        54        59
  vocati |        50     46.76  9.318754        40        46        54
---------+--------------------------------------------------------------
   Total |       200    52.775  9.478586      45.5        54        60
-------------------------------------------------------------------------
```

The complete code for the concepts that are described above is as follows:

```
capture log close
log using example1.txt, text replace

use http://www.ats.ucla.edu/stat/data/hs0, clear
describe
codebook

summarize read math science write
tabstat read write math, by(prgtype) stat(n mean sd)
tabstat write, by(prgtype) stat(n mean sd p25 p50 p75)

log close
```