```
*******************************
* Penguin Analysis
* Demonstration Do File
*******************************

* So many projects have the same, or similar, workflow.

* have a question ->
* get data ->
* process and clean data ->
* analyze data ->
* visualize data ->
* make conclusions

/* do files are useful to preserve
a record of your work. They help
to keep an audit trail of the
decisions that you have made. */

/* do files thus serve as a way of creating an
automated, replicable and documented workflow
as well as finding and minimizing errors */

* A `*` character at the beginning of a line makes that
line a comment

/* You can also use asterisk slash to denote multiple
lines of comment */

*******************************
* get data
*******************************

* a good workflow habit is to
* always--or at least frequently--
* work from your raw data.

* i.e. run your script so you are always--
* or at least often--opening your raw data,
* cleaning the data, creating new variables,
* and then running analyses.

```

```
41  clear all // clear the workspace
42
43  * get data from web
44
45  use "penguins.dta", clear
46
47  ********************************
48  * take a look at the data
49  ********************************
50
51  * NB if you have a lot of variables, the commands below
    will produce a lot of (too much) output
52
53  * you may need to `describe` or `codebook` specific
    variables
54
55  describe // describe the variables
56
57  codebook // full descriptions of all the variables;
    produces a lot of output
58
59  ********************************
60  * descriptive statistics
61  ********************************
62
63  summarize // descriptive statistics for all variables
64
65  summarize body_mass_g // descriptive statistics for this
    variable
66
67  tabulate species // tabulate this categorical variable
68
69  * dtable is a useful new command
70  * for producing tables of descriptive statistics
71  * be sure to denote indicator variables with an `i.`
72
73  dtable culmen_length_mm body_mass_g i.species
74
75  ********************************
76  * data wrangling
77  ********************************
78
```

```
* find variables of interest

lookfor mass // look for a variable w a particular keyword

* sometimes it is useful to `keep` only the variables in which you have an interest
* to reduce the size of the data set

* recode variables

generate big_penguin = body_mass_g > 4000 // create a big penguin variable

tabulate big_penguin

*********************************
* ANOVA
*********************************

oneway body_mass_g species, tabulate

*********************************
* regression
*********************************

regress culmen_length_mm body_mass_g

est store M1 // store regression estimates

regress culmen_length_mm body_mass_g i.species

est store M2 // store regression estimates

* /// indicates that a command spans multiple lines

etable, estimates(M1 M2) /// nicely formatted table of regression estimates
cstat(_r_b) /// beta's only
showstars showstarsnote // show stars and note

*********************************
* graph
```

```
118 ******************************
119
120 graph bar body_mass_g, over(species) // bar graph
121
122 twoway scatter culmen_length_mm body_mass_g // scatterplot
123
124
125
126
127
128
```