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

* So many projects have the same, or similar, workflow.
* DO YOUR THINKING IN CODE!!!

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

```
79
80  * find variables of interest
81
82  lookfor mass // look for a variable w a particular keyword
83
84  * sometimes it is useful to `keep` only the variables in
    which you have an interest
85  * to reduce the size of the data set
86
87  * recode variables
88
89  generate big_penguin = body_mass_g > 4000 // create a
    big penguin variable
90
91  tabulate big_penguin
92
93  ********************************
94  * ANOVA
95  ********************************
96
97  oneway body_mass_g species, tabulate
98
99  ********************************
100 * regression
101 ********************************
102
103 regress culmen_length_mm body_mass_g
104
105 est store M1 // store regression estimates
106
107 regress culmen_length_mm body_mass_g i.species
108
109 est store M2 // store regression estimates
110
111 * /// indicates that a command spans multiple lines
112
113 etable, estimates(M1 M2) /// nicely formatted table of
    regression estimates
114 cstat(_r_b) /// beta's only
115 showstars showstarsnote // show stars and note
116
117 ********************************
```

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

graph bar body_mass_g, over(species) // bar graph

twoway scatter culmen_length_mm body_mass_g // scatterplot
```