```
1   ******************************
2   * Penguin Analysis
3   * Demonstration Do File
4   ******************************
5
6   * So many projects have the same, or similar, workflow.
7
8   * have a question ->
9   * get data ->
10  * process and clean data ->
11  * analyze data ->
12  * visualize data ->
13  * make conclusions
14
15  /* do files are useful to preserve
16  a record of your work. They help
17  to keep an audit trail of the
18  decisions that you have made. */
19
20  /* do files thus serve as a way of creating an
21  automated, replicable and documented workflow
22  as well as finding and minimizing errors */
23
24  * A `*` character at the beginning of a line makes that line a comment
25
26  /* You can also use asterisk slash to denote multiple lines of
    comment */
27
28  ******************************
29  * get data
30  ******************************
31
32  * a good workflow habit is to
33  * always--or at least frequently--
34  * work from your raw data.
35
36  * i.e. run your script so you are always--
37  * or at least often--opening your raw data,
38  * cleaning the data, creating new variables,
39  * and then running analyses.
40
41  clear all // clear the workspace
42
43  * get data from web
44
45  use
```

```
    "https://github.com/agrogan1/Stata/raw/master/do-files/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  ******************************
70  * data wrangling
71  ******************************
72
73  * find variables of interest
74
75  lookfor mass // look for a variable w a particular keyword
76
77  * sometimes it is useful to `keep` only the variables in which you
    have an interest
78  * to reduce the size of the data set
79
80  * recode variables
81
82  generate big_penguin = body_mass_g > 4000 // create a big penguin
    variable
83
84  tabulate big_penguin
85
```

```
86  ******************************
87  * ANOVA
88  ******************************
89
90  oneway body_mass_g species, tabulate
91
92  ******************************
93  * regression
94  ******************************
95
96  regress culmen_length_mm body_mass_g
97
98  est store M1 // store regression estimates
99
100 regress culmen_length_mm body_mass_g i.species
101
102 est store M2 // store regression estimates
103
104 est table M1 M2, b(%7.4f) star // nicely formatted table of
    regression estimates
105
106 ******************************
107 * graph
108 ******************************
109
110 graph bar body_mass_g, over(species) scheme(s1color) // bar graph
111
112 twoway scatter culmen_length_mm body_mass_g, scheme(s1color) //
    scatterplot
113
114
115
116
117
118
```