Getting to Know Your Data Set in Stata 12

Table of Contents

This is a guideline for Stata 12. Updated versions of Stata may have differing commands.

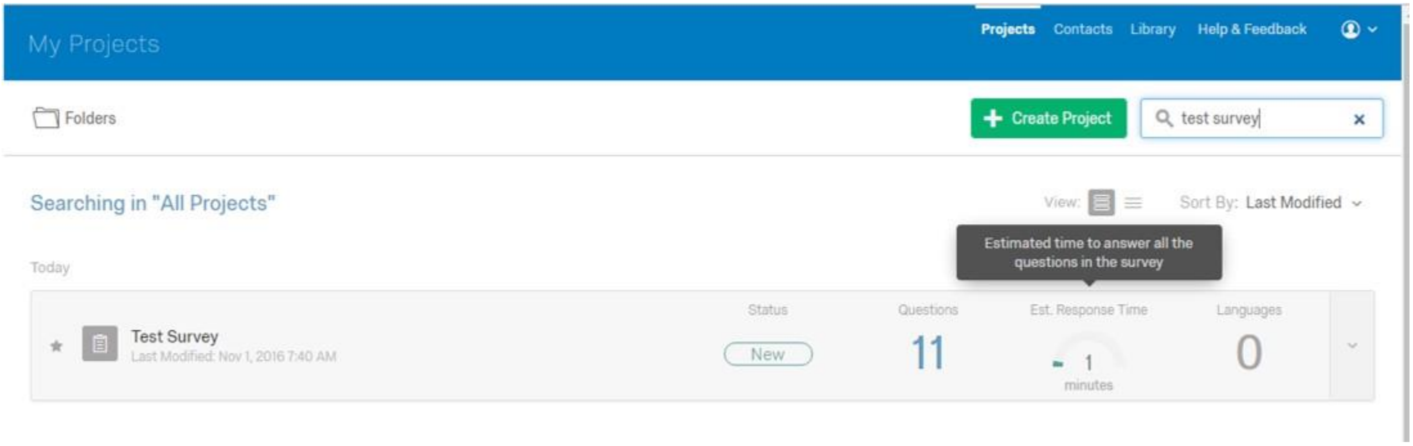
I. Data Set

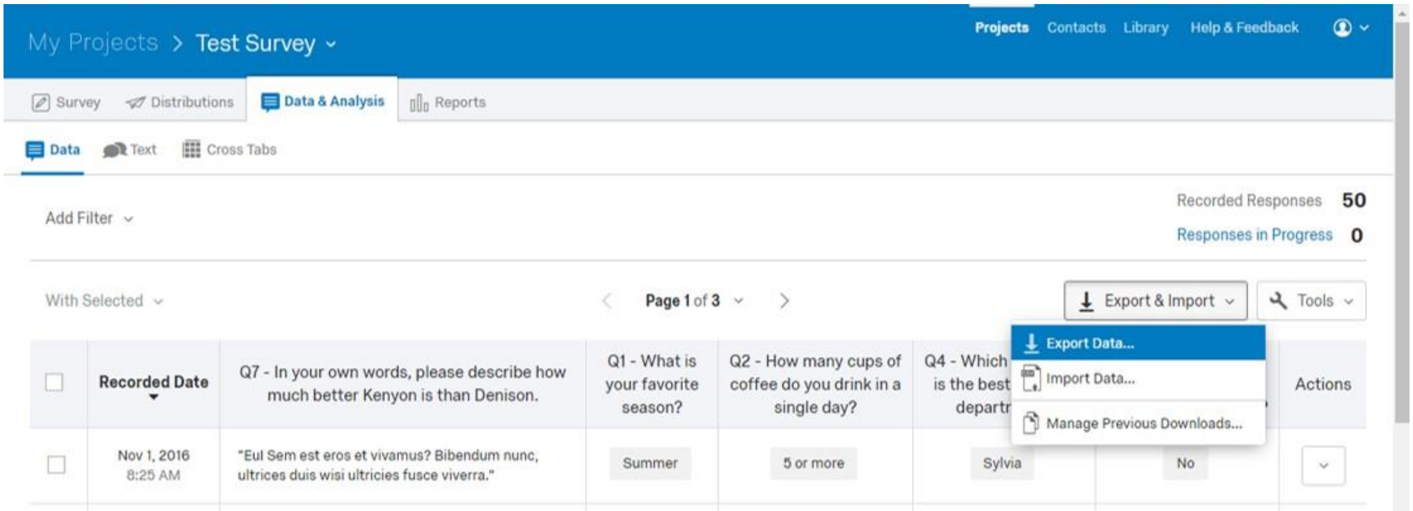
For our data analysis, we will use a fake data set entitled **Test Survey**. The variables are the text and numerical values (these will have **num** in front of the variable name) for favorite season (**season** and **numseason**), number of cups of coffee consumed each day (**coffee** and **numcoffee**), favorite restaurant in a local town (**restaurant** and **numrestaurant**), favorite computer's name (**computer** and **numcomputer**), favorite meal in the dining hall (**peircemeals** and **numpeircemeals**), a question pertaining to the fate of the characters on the popular television show *Lost* (**lost** and **numlost**), how much they like Kenyon with respect to another school (**kenyon** and **numkenyon**), students favorite dessert (**dessert** and **numdessert**), whether they watch a particular television show (**AHS** and **numAHS**), what their favorite season of said television show is (**favoriteAHS** and **numfavoriteAHS**), and students' favorite building on campus (**building** and **numbuilding**). This survey has 50 observations. We are going to use these variables to examine some simple summary statistics and a simple regression. We will use the names of our variables. If you were to replicate this analysis, use your variable names rather than our variable names.

II. Acquiring your Data Set from Qualtrics

Most of your data sets will come from the survey collection website Qualtrics. To get your desired data set, go to kenyon.qualtrics.com.

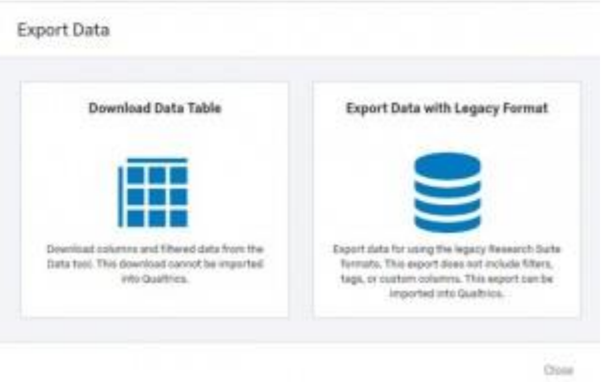
1. To find your desired data, search the name of the survey in the search bar in the top right corner (it will be next to "Create Project"). Once you have found your survey, click "Data & Analysis then select "Export & Import" and choose "Export Data" in order to acquire the data set.





□	Recorded Date	Q7 - In your own words, please describe how much better Kenyon is than Denison.	Q1 - What is your favorite season?	Q2 - How many cups of coffee do you drink in a single day?	Q4 - Which is the best departr		Actions
□	Nov 1, 2016 8:25 AM	"Eul Sem est eros et vivamus? Bibendum nunc, ultrices duis wisi ultricies fusce viverra."	Summer	5 or more	Sylvia	No	

2. Upon choosing "Export Data", you will then want to select "Download Data Table". Several options will appear as to which format you can download your data as. This is up to you, but we chose "CSV" so that we could keep the choice text. We then opened this data in Excel, and copy and pasted it into Stata's Data Editor.

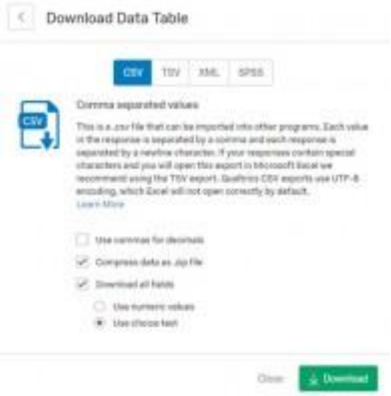

Export Data

Download Data Table

Download columns and filtered data from the Data tool. This download cannot be imported into Qualtrics.

Export Data with Legacy Format

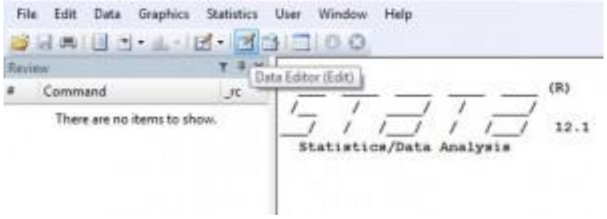
Export data for using the legacy Research Suite formats. This export does not include filters, tags, or custom columns. This export can be imported into Qualtrics.


Download Data Table

Comma separated values

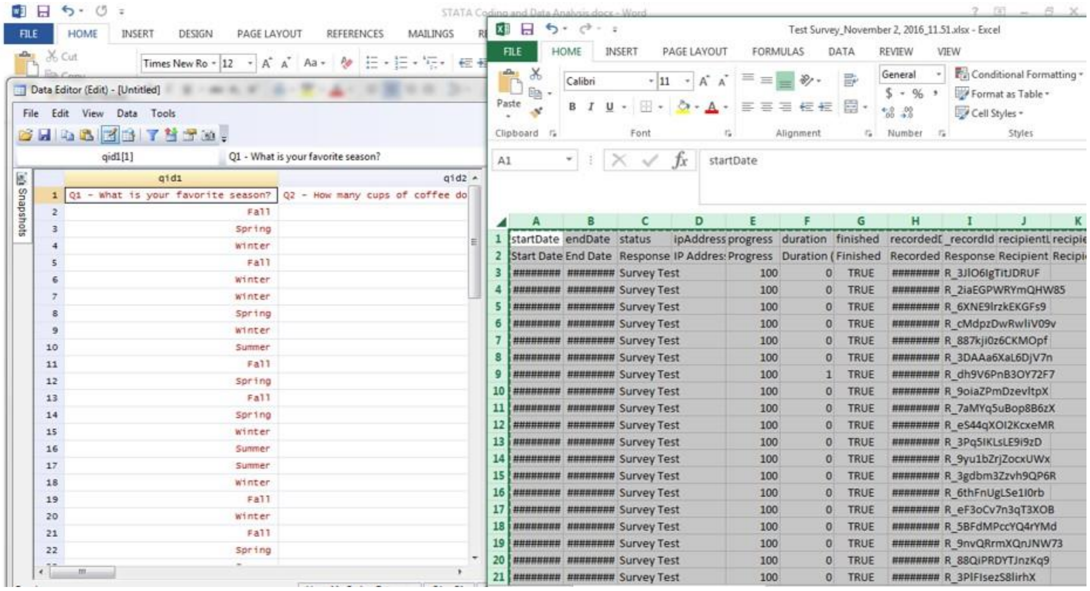
This is a .csv file that can be imported into other programs. Each value in the response is separated by a comma and each response is separated by a newline character. If your responses contain special characters and you will open this export in Microsoft Excel we recommend using the TSV export. Qualtrics CSV exports use UTF-8 encoding, which Excel will not open correctly by default.
Learn More

□ Use commas for decimals
☑ Compress data as .zip file
☑ Download all fields
○ Use numeric values
● Use choice text



There are no items to show.

Statistics/Data Analysis 12.1



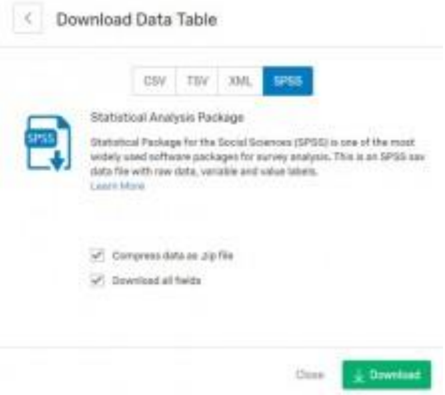
qid1[1] Q1 - What is your favorite season?

	qid1	qid2
1	Q1 - What is your favorite season?	Q2 - How many cups of coffee do
2	Fall	
3	Spring	
4	Winter	
5	Fall	
6	Winter	
7	Winter	
8	Spring	
9	Winter	
10	Summer	
11	Fall	
12	Spring	
13	Fall	
14	Spring	
15	Winter	
16	Summer	
17	Summer	
18	Winter	
19	Fall	
20	Winter	
21	Fall	
22	Spring	

A1 startDate

	A	B	C	D	E	F	G	H	I	J	K
1	startDate	endDate	status	ipAddress	progress	duration	finished	recordedD	_recordId	recipientL	recipie
2	Start Date	End Date	Response	IP Address	Progress	Duration (	Finished	Recorded	Response	Recipient	Recipi
3	########	########	Survey Test		100	0	TRUE	########	R_3JlO6IgTitJDRUF		
4	########	########	Survey Test		100	0	TRUE	########	R_2iaEGPWRYmQHW85		
5	########	########	Survey Test		100	0	TRUE	########	R_6XNE9lrzkEKGFs9		
6	########	########	Survey Test		100	0	TRUE	########	R_cMdpzDwRwliV09v		
7	########	########	Survey Test		100	0	TRUE	########	R_887kji0z6CKMOpf		
8	########	########	Survey Test		100	0	TRUE	########	R_3DAAa6XaL6DjV7n		
9	########	########	Survey Test		100	1	TRUE	########	R_dh9V6PnB3OY72F7		
10	########	########	Survey Test		100	0	TRUE	########	R_9oiaZPmDzevltpX		
11	########	########	Survey Test		100	0	TRUE	########	R_7aMYq5uBop8B6zX		
12	########	########	Survey Test		100	0	TRUE	########	R_eS44qXOI2KcxeMR		
13	########	########	Survey Test		100	0	TRUE	########	R_3Pq5IKLsLE9i9zD		
14	########	########	Survey Test		100	0	TRUE	########	R_9yu1bZrjZocxUWx		
15	########	########	Survey Test		100	0	TRUE	########	R_3gdbm3Zzvh9QP6R		
16	########	########	Survey Test		100	0	TRUE	########	R_6thFnUgLSe1I0rb		
17	########	########	Survey Test		100	0	TRUE	########	R_eF3oCv7n3qT3XOB		
18	########	########	Survey Test		100	0	TRUE	########	R_5BFdMPccYQ4rYMd		
19	########	########	Survey Test		100	0	TRUE	########	R_9nvQRrmXQnJNW73		
20	########	########	Survey Test		100	0	TRUE	########	R_88QiPRDYTJnzKq9		
21	########	########	Survey Test		100	0	TRUE	########	R_3PlFIsezS8lirhX		

3. To get the numeric data, we again download the data from Qualtrics, but this time choose "SPSS". Then you will go to the SPSS statistical package and copy the desired numerical values from its "Data View" tab and paste them into Stata's Data Editor.



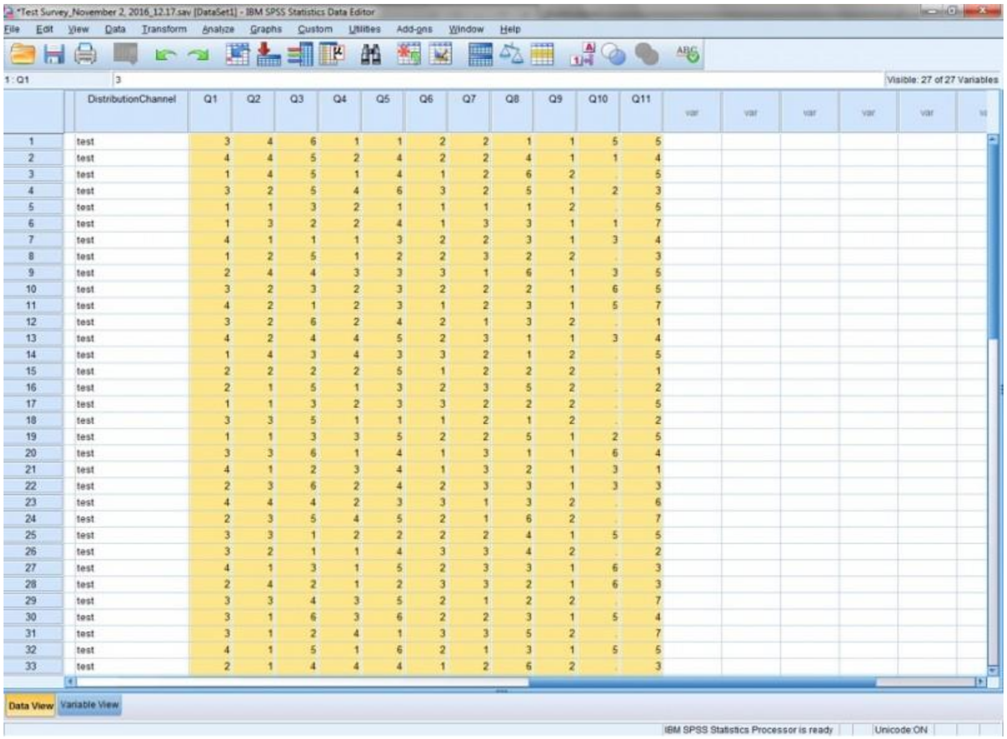
Statistical Analysis Package

Statistical Package for the Social Sciences (SPSS) is one of the most widely used software packages for survey analysis. This is an SPSS sav data file with raw data, variable and value labels.

Learn More

☑ Compress data as .zip file

☑ Download all fields

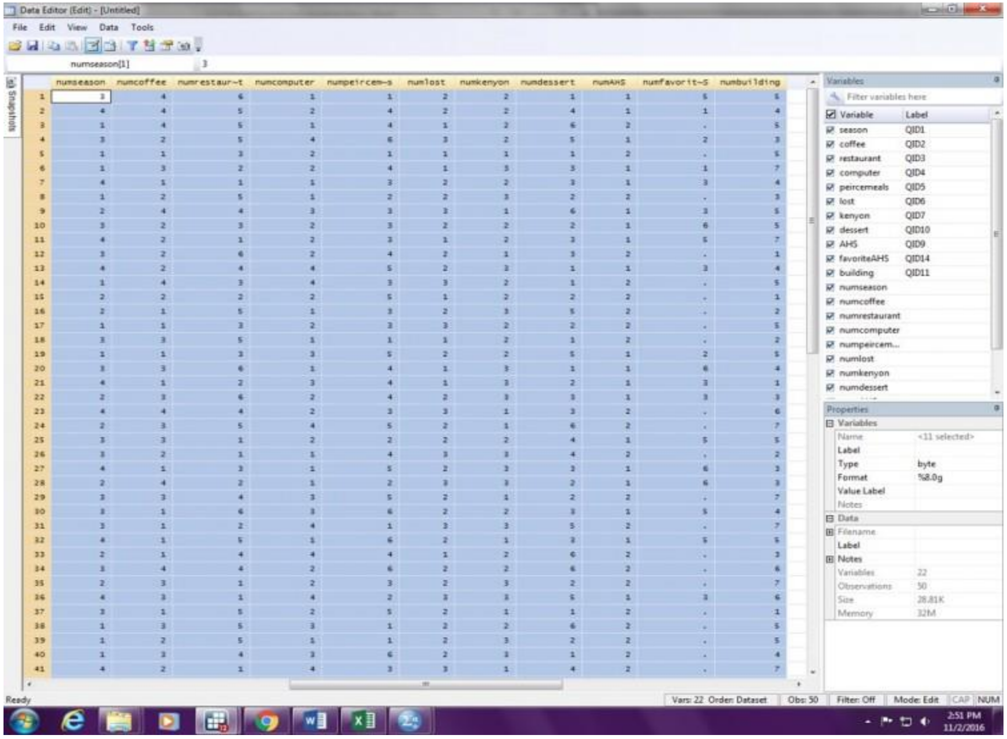

*Test Survey_November 2, 2016_12.17.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Custom Utilities Add-ons Window Help
Data View
Variable View

1 : Q1 3

Visible: 27 of 27 Variables

	DistributionChannel	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11
1	test	3	4	6	1	1	2	2	1	1	5	5
2	test	4	4	5	2	4	2	2	4	1	1	4
3	test	1	4	5	1	4	1	2	6	2	.	5
4	test	3	2	5	4	6	3	2	5	1	2	3
5	test	1	1	3	2	1	1	1	1	2	.	5
6	test	1	3	2	2	4	1	3	3	1	1	7
7	test	4	1	1	1	3	2	2	3	1	3	4
8	test	1	2	5	1	2	2	3	2	2	.	3
9	test	2	4	4	3	3	3	1	6	1	3	5
10	test	3	2	3	2	3	2	2	2	1	6	5
11	test	4	2	1	2	3	1	2	3	1	5	7
12	test	3	2	6	2	4	2	1	3	2	.	1
13	test	4	2	4	4	5	2	3	1	1	3	4
14	test	1	4	3	4	3	3	2	1	2	.	5
15	test	2	2	2	2	5	1	2	2	2	.	1
16	test	2	1	5	1	3	2	3	5	2	.	2
17	test	1	1	3	2	3	3	2	2	2	.	5
18	test	3	3	5	1	1	1	2	1	2	.	2
19	test	1	1	3	3	5	2	2	5	1	2	5
20	test	3	3	6	1	4	1	3	1	1	6	4
21	test	4	1	2	3	4	1	3	2	1	3	1
22	test	2	3	6	2	4	2	3	3	1	3	3
23	test	4	4	4	2	3	3	1	3	2	.	6
24	test	2	3	5	4	5	2	1	6	2	.	7
25	test	3	3	1	2	2	2	2	4	1	5	5
26	test	3	2	1	1	4	3	3	4	2	.	2
27	test	4	1	3	1	5	2	3	3	1	6	3
28	test	2	4	2	1	2	3	3	2	1	6	3
29	test	3	3	4	3	5	2	1	2	2	.	7
30	test	3	1	6	3	6	2	2	3	1	5	4
31	test	3	1	2	4	1	3	3	5	2	.	7
32	test	4	1	5	1	6	2	1	3	1	5	5
33	test	2	1	4	4	4	1	2	6	2	.	3

IBM SPSS Statistics Processor is ready | Unicode:ON



numseason[1] 3

	numseason	numcoffee	numrestaur~t	numcomputer	numpeircem~s	numlost	numkenyon	numdessert	numAHS	numfavorit~S	numbuilding
1	3	4	6	1	1	2	2	1	1	5	5
2	4	4	5	2	4	2	2	4	1	1	4
3	1	4	5	1	4	1	2	6	2	.	5
4	3	2	5	4	6	3	2	5	1	2	3
5	1	1	3	2	1	1	1	1	2	.	5
6	1	3	2	2	4	1	3	3	1	1	7
7	4	1	1	1	3	2	2	3	1	3	4
8	1	2	5	1	2	2	3	2	2	.	3
9	2	4	4	3	3	3	1	6	1	3	5
10	3	2	3	2	3	2	2	2	1	6	5
11	4	2	1	2	3	1	2	3	1	5	7
12	3	2	6	2	4	2	1	3	2	.	1
13	4	2	4	4	5	2	3	1	1	3	4
14	1	4	3	4	3	3	2	1	2	.	5
15	2	2	2	2	5	1	2	2	2	.	1
16	2	1	5	1	3	2	3	5	2	.	2
17	1	1	3	2	3	3	2	2	2	.	5
18	3	3	5	1	1	1	2	1	2	.	2
19	1	1	3	3	5	2	2	5	1	2	5
20	3	3	6	1	4	1	3	1	1	6	4
21	4	1	2	3	4	1	3	2	1	3	1
22	2	3	6	2	4	2	3	3	1	3	3
23	4	4	4	2	3	3	1	3	2	.	6
24	2	3	5	4	5	2	1	6	2	.	7
25	3	3	1	2	2	2	2	4	1	5	5
26	3	2	1	1	4	3	3	4	2	.	2
27	4	1	3	1	5	2	3	3	1	6	3
28	2	4	2	1	2	3	3	2	1	6	3
29	3	3	4	3	5	2	1	2	2	.	7
30	3	1	6	3	6	2	2	3	1	5	4
31	3	1	2	4	1	3	3	5	2	.	7
32	4	1	5	1	6	2	1	3	1	5	5
33	2	1	4	4	4	1	2	6	2	.	3
34	3	4	4	2	6	2	2	6	2	.	6
35	2	3	1	2	3	2	3	2	2	.	7
36	4	3	1	4	2	3	3	5	1	3	6
37	3	1	5	2	5	2	1	1	2	.	1
38	1	3	5	3	1	2	2	6	2	.	5
39	1	2	5	1	1	2	3	2	2	.	5
40	1	3	4	3	6	2	1	1	2	.	4
41	4	2	1	4	3	3	1	4	2	.	7

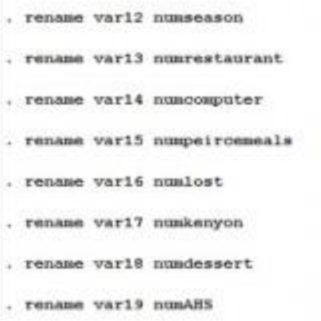
☑ Variable	Label
☑ season	QID1
☑ coffee	QID2
☑ restaurant	QID3
☑ computer	QID4
☑ peircemeals	QID5
☑ lost	QID6
☑ kenyon	QID7
☑ dessert	QID10
☑ AHS	QID9
☑ favoriteAHS	QID14
☑ building	QID11
☑ numseason	
☑ numcoffee	
☑ numrestaurant	
☑ numcomputer	
☑ numpeircem...	
☑ numlost	
☑ numkenyon	
☑ numdessert	

⊟ Variables	
Name	<11 selected>
Label	
Type	byte
Format	%8.0g
Value Label	
Notes	
⊟ Data	
⊞ Filename	
Label	
⊞ Notes	
Variables	22
Observations	50
Size	28.81K
Memory	32M

Ready | Vars: 22 Order: Dataset | Obs: 50 | Filter: Off | Mode: Edit | CAP | NUM

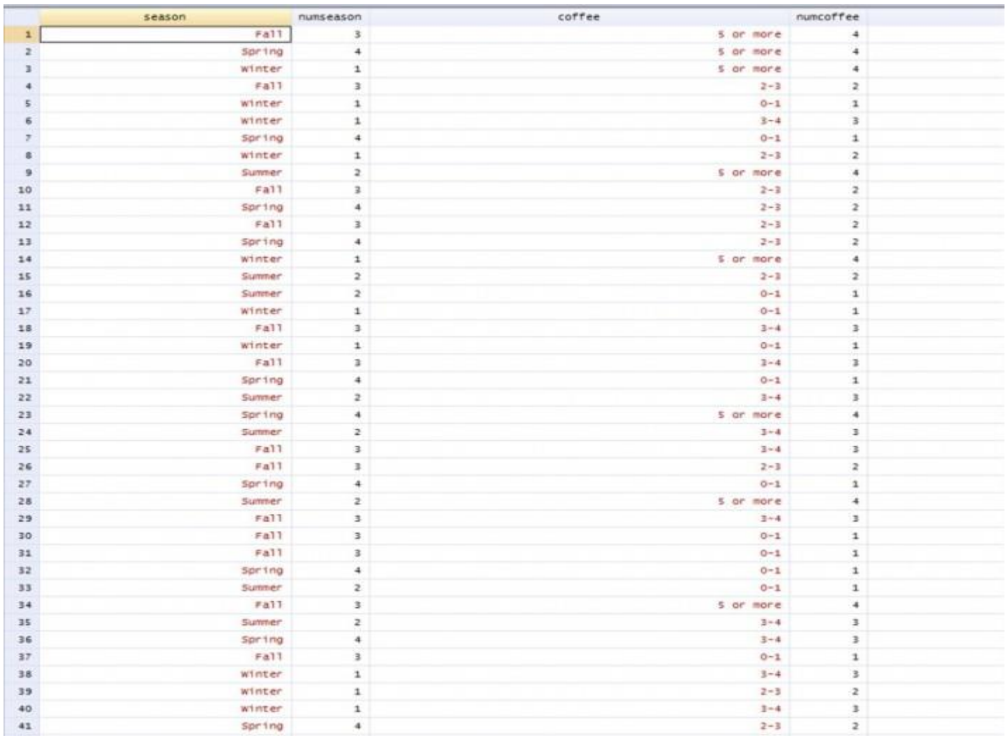
2:51 PM
11/2/2016

4. These numerical values will be named "var1", "var2", etc. To help make more sense of the data, use “rename var12 numseason”.



```
. rename var12 numseason

. rename var13 numrestaurant

. rename var14 numcomputer

. rename var15 numpeiromeals

. rename var16 numlost

. rename var17 numkenyon

. rename var18 numdessert

. rename var19 numAHS
```

5. Finally, reorder your data set so that choice text is next to the numerical values associated with each variable. To do this, use “order season numseason coffee numcoffee…building numbuilding”.

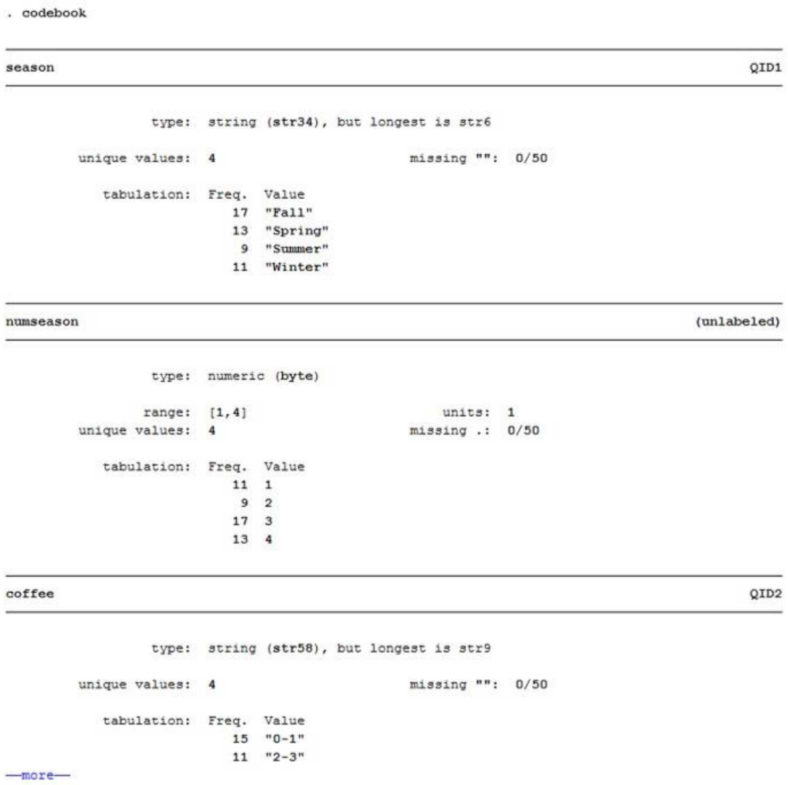


	season	numseason	coffee	numcoffee
1	Fall	3	5 or more	4
2	Spring	4	5 or more	4
3	winter	1	5 or more	4
4	Fall	3	2-3	2
5	winter	1	0-1	1
6	winter	1	3-4	3
7	Spring	4	0-1	1
8	winter	1	2-3	2
9	Summer	2	5 or more	4
10	Fall	3	2-3	2
11	Spring	4	2-3	2
12	Fall	3	2-3	2
13	Spring	4	2-3	2
14	winter	1	5 or more	4
15	Summer	2	2-3	2
16	Summer	2	0-1	1
17	winter	1	0-1	1
18	Fall	3	3-4	3
19	winter	1	0-1	1
20	Fall	3	3-4	3
21	Spring	4	0-1	1
22	Summer	2	3-4	3
23	Spring	4	5 or more	4
24	Summer	2	3-4	3
25	Fall	3	3-4	3
26	Fall	3	2-3	2
27	Spring	4	0-1	1
28	Summer	2	5 or more	4
29	Fall	3	3-4	3
30	Fall	3	0-1	1
31	Fall	3	0-1	1
32	Spring	4	0-1	1
33	Summer	2	0-1	1
34	Fall	3	5 or more	4
35	Summer	2	3-4	3
36	Spring	4	3-4	3
37	Fall	3	0-1	1
38	winter	1	3-4	3
39	winter	1	2-3	2
40	winter	1	3-4	3
41	Spring	4	2-3	2

III. Stata Output

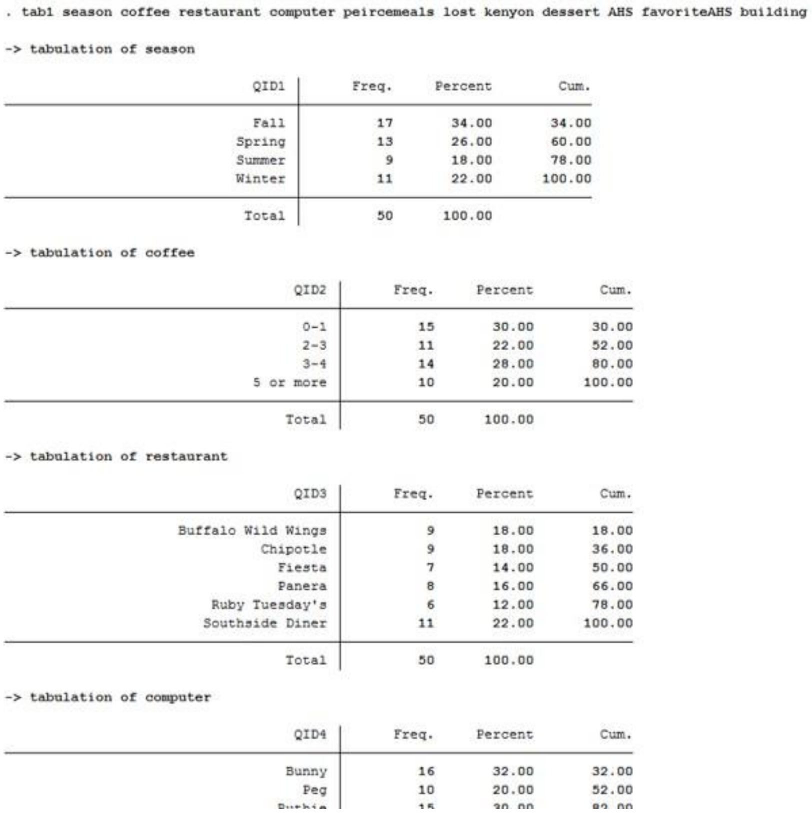
Getting to know your data set:

1. For a quick overview of your data set, use the command “codebook” to get the total number of observations for each variable and the frequency for each possible response.

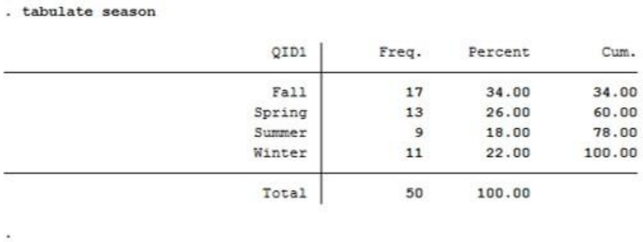


```
. codebook

-------------------------------------------------------------------------------
season                                                                     QID1
-------------------------------------------------------------------------------

                  type:  string (str34), but longest is str6

         unique values:  4                        missing "":  0/50

            tabulation:  Freq.  Value
                            17  "Fall"
                            13  "Spring"
                             9  "Summer"
                            11  "Winter"

-------------------------------------------------------------------------------
numseason                                                           (unlabeled)
-------------------------------------------------------------------------------

                  type:  numeric (byte)

                 range:  [1,4]                        units:  1
         unique values:  4                        missing .:  0/50

            tabulation:  Freq.  Value
                            11  1
                             9  2
                            17  3
                            13  4

-------------------------------------------------------------------------------
coffee                                                                     QID2
-------------------------------------------------------------------------------

                  type:  string (str58), but longest is str9

         unique values:  4                        missing "":  0/50

            tabulation:  Freq.  Value
                            15  "0-1"
                            11  "2-3"
--more--
```

2. Use the “tab1 season coffee...building” command to get the frequency and percent of each response for each variable. You could also use the "tabulate season" command to just get the frequency and percent for one variable and its responses.

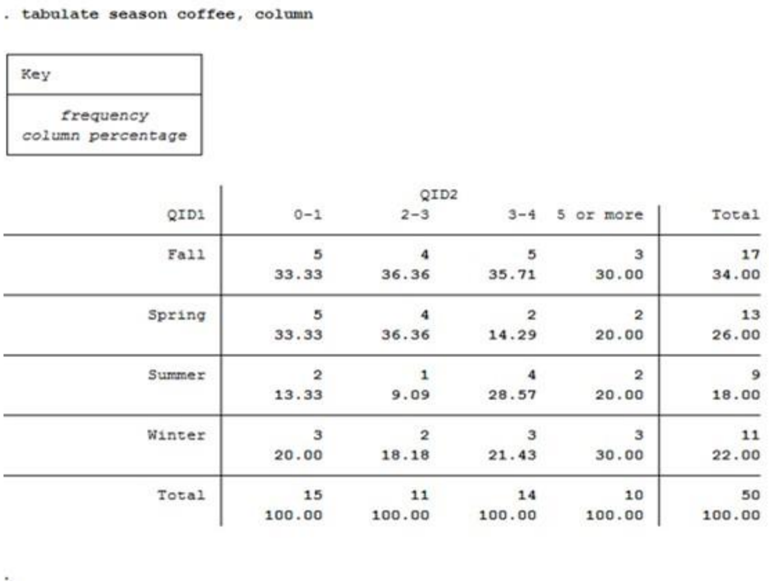


```
. tab1 season coffee restaurant computer peircemeals lost kenyon dessert AHS favoriteAHS building

-> tabulation of season

               QID1 |      Freq.     Percent        Cum.
--------------------+-----------------------------------
               Fall |         17       34.00       34.00
             Spring |         13       26.00       60.00
             Summer |          9       18.00       78.00
             Winter |         11       22.00      100.00
--------------------+-----------------------------------
              Total |         50      100.00

-> tabulation of coffee

               QID2 |      Freq.     Percent        Cum.
--------------------+-----------------------------------
                0-1 |         15       30.00       30.00
                2-3 |         11       22.00       52.00
                3-4 |         14       28.00       80.00
          5 or more |         10       20.00      100.00
--------------------+-----------------------------------
              Total |         50      100.00

-> tabulation of restaurant

               QID3 |      Freq.     Percent        Cum.
--------------------+-----------------------------------
  Buffalo Wild Wings |          9       18.00       18.00
            Chipotle |          9       18.00       36.00
              Fiesta |          7       14.00       50.00
              Panera |          8       16.00       66.00
      Ruby Tuesday's |          6       12.00       78.00
    Southside Diner |         11       22.00      100.00
--------------------+-----------------------------------
              Total |         50      100.00

-> tabulation of computer

               QID4 |      Freq.     Percent        Cum.
--------------------+-----------------------------------
              Bunny |         16       32.00       32.00
                Peg |         10       20.00       52.00
             Ruthie |         15       30.00       82.00
```



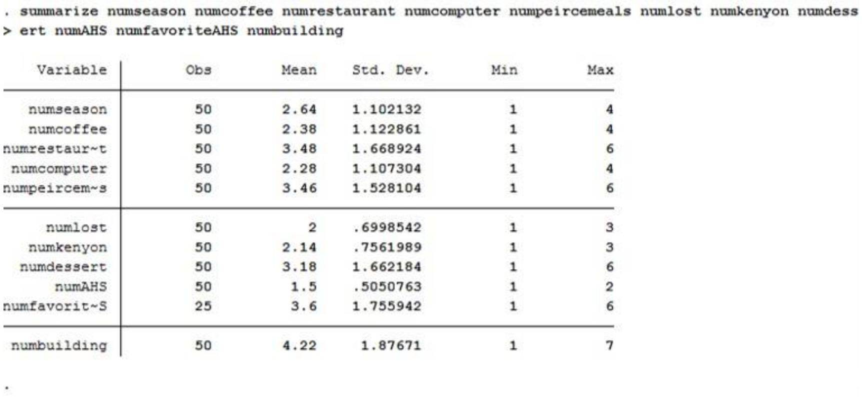
```
. tabulate season

               QID1 |      Freq.     Percent        Cum.
--------------------+-----------------------------------
               Fall |         17       34.00       34.00
             Spring |         13       26.00       60.00
             Summer |          9       18.00       78.00
             Winter |         11       22.00      100.00
--------------------+-----------------------------------
              Total |         50      100.00

.
```

3. To compare different variables to one another, use the command "tabulate season coffee, column". This command will give you the frequency and percentage of a data point fulfilling two different criteria (i.e. how many responses answered "Fall" and "0-1" cups of coffee and so forth).



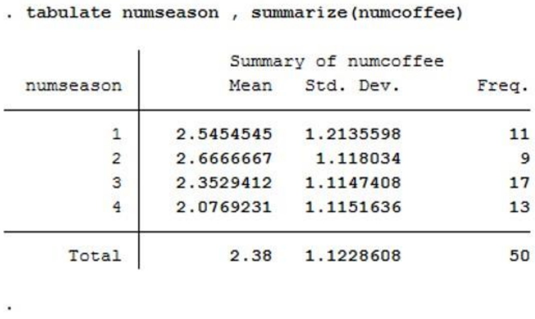
```
. tabulate season coffee, column

+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

           |                  QID2
      QID1 |       0-1        2-3        3-4  5 or more |     Total
-----------+--------------------------------------------+----------
      Fall |         5          4          5          3 |        17 
           |     33.33      36.36      35.71      30.00 |     34.00 
-----------+--------------------------------------------+----------
    Spring |         5          4          2          2 |        13 
           |     33.33      36.36      14.29      20.00 |     26.00 
-----------+--------------------------------------------+----------
    Summer |         2          1          4          2 |         9 
           |     13.33       9.09      28.57      20.00 |     18.00 
-----------+--------------------------------------------+----------
    Winter |         3          2          3          3 |        11 
           |     20.00      18.18      21.43      30.00 |     22.00 
-----------+--------------------------------------------+----------
     Total |        15         11         14         10 |        50 
           |    100.00     100.00     100.00     100.00 |    100.00 

.
```

4. To acquire summary statistics regarding the variables, you will have to use the numerical values. The command needed would be “summarize numseason numcoffee numrestaurant numcomputer numpeircemeals numlost numkenyon numdessert numAHS numfavoriteAHS numbuilding”. This will give you summary statistics such as mean, number of observations, standard deviation, minimums, and maximums.



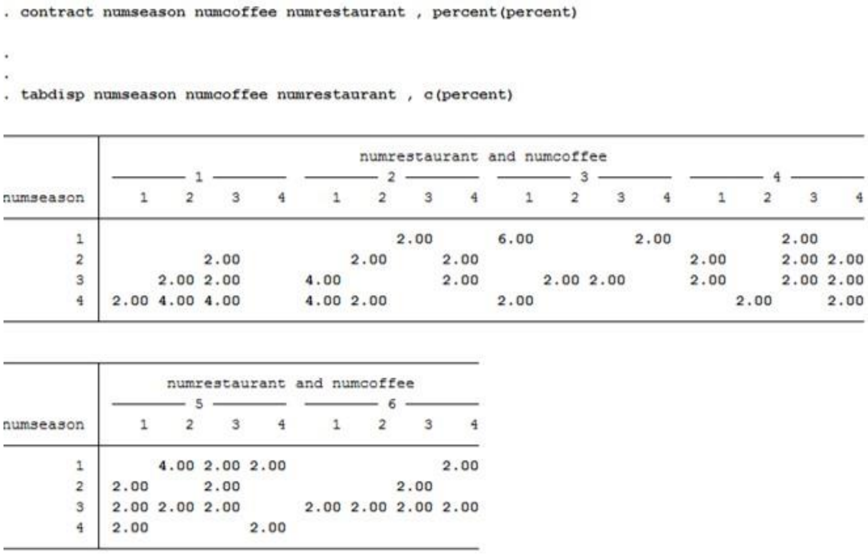
```
. summarize numseason numcoffee numrestaurant numcomputer numpeircemeals numlost numkenyon numdess
> ert numAHS numfavoriteAHS numbuilding

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
   numseason |        50        2.64    1.102132          1          4
   numcoffee |        50        2.38    1.122861          1          4
numrestaur~t |        50        3.48    1.668924          1          6
 numcomputer |        50        2.28    1.107304          1          4
numpeircem~s |        50        3.46    1.528104          1          6
-------------+--------------------------------------------------------
     numlost |        50           2    .6998542          1          3
   numkenyon |        50        2.14    .7561989          1          3
  numdessert |        50        3.18    1.662184          1          6
      numAHS |        50         1.5    .5050763          1          2
numfavorit~S |        25         3.6    1.755942          1          6
-------------+--------------------------------------------------------
 numbuilding |        50        4.22     1.87671          1          7

.
```

5. Use the "tabulate numseason, summarize(numcoffee)” to get summary statistics for a variable with respect to another variable.



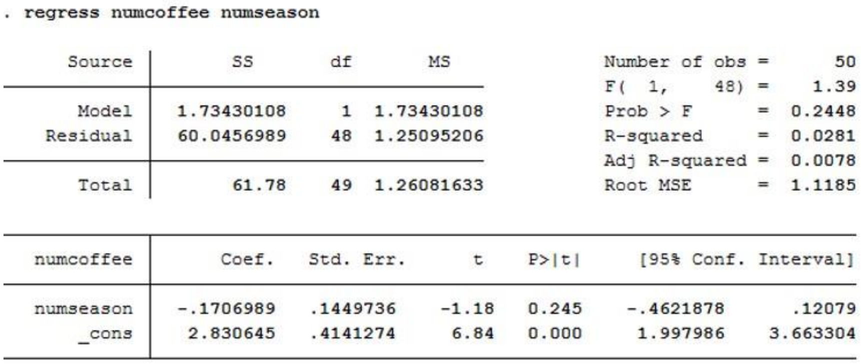
```
. tabulate numseason , summarize(numcoffee)

            |       Summary of numcoffee
  numseason |        Mean   Std. Dev.       Freq.
------------+------------------------------------
          1 |   2.5454545   1.2135598          11
          2 |   2.6666667    1.118034           9
          3 |   2.3529412   1.1147408          17
          4 |   2.0769231   1.1151636          13
------------+------------------------------------
      Total |        2.38   1.1228608          50

.
```

6. In order to compare three variables to each other, use the commands “contract numseason numcoffee numrestaurant, percent(percent)” and then “tabdisp numseason numcoffee numrestaurant, c(percent)”. The first command will contract the other variables so that only the ones of interest will be displayed. The second command will then create a display table that shows the number of responses that were in each combination of the three different variable's responses.



```
. contract numseason numcoffee numrestaurant , percent(percent)

.
.
. tabdisp numseason numcoffee numrestaurant , c(percent)

----------+-----------------------------------------------------------------------------------
          |                            numrestaurant and numcoffee
          | -------- 1 --------  -------- 2 --------  -------- 3 --------  -------- 4 --------
numseason |    1    2    3    4     1    2    3    4     1    2    3    4     1    2    3    4
----------+-----------------------------------------------------------------------------------
        1 |                                2.00       6.00           2.00            2.00
        2 |           2.00            2.00      2.00                       2.00      2.00 2.00
        3 |      2.00 2.00       4.00           2.00       2.00 2.00       2.00      2.00 2.00
        4 | 2.00 4.00 4.00       4.00 2.00            2.00                      2.00      2.00
----------------------------------------------------------------------------------------------

----------+----------------------------------------
          |       numrestaurant and numcoffee
          | -------- 5 --------  -------- 6 --------
numseason |    1    2    3    4     1    2    3    4
----------+----------------------------------------
        1 |      4.00 2.00 2.00                 2.00
        2 | 2.00      2.00                 2.00
        3 | 2.00 2.00 2.00       2.00 2.00 2.00 2.00
        4 | 2.00           2.00
---------------------------------------------------
```

7. Finally, to run a regression to examine the causal relationships between two or more variables, use the command “regress numcoffee numseason”, taking care to put the dependent variable first and the explanatory variable(s) after that. In other words, put the variable that you are trying to predict first and the variable(s) that help you predict this variable after.



```
. regress numcoffee numseason

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  1,    48) =    1.39
       Model |  1.73430108     1  1.73430108           Prob > F      =  0.2448
    Residual |  60.0456989    48  1.25095206           R-squared     =  0.0281
-------------+------------------------------           Adj R-squared =  0.0078
       Total |       61.78    49  1.26081633           Root MSE      =  1.1185

------------------------------------------------------------------------------
   numcoffee |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   numseason |  -.1706989   .1449736    -1.18   0.245    -.4621878      .12079
       _cons |   2.830645   .4141274     6.84   0.000     1.997986    3.663304
------------------------------------------------------------------------------
```