Getting to Know Your Data Set in RStudio

Table of Contents

Please note that these guidelines were written using RStudio Version 1.0.44

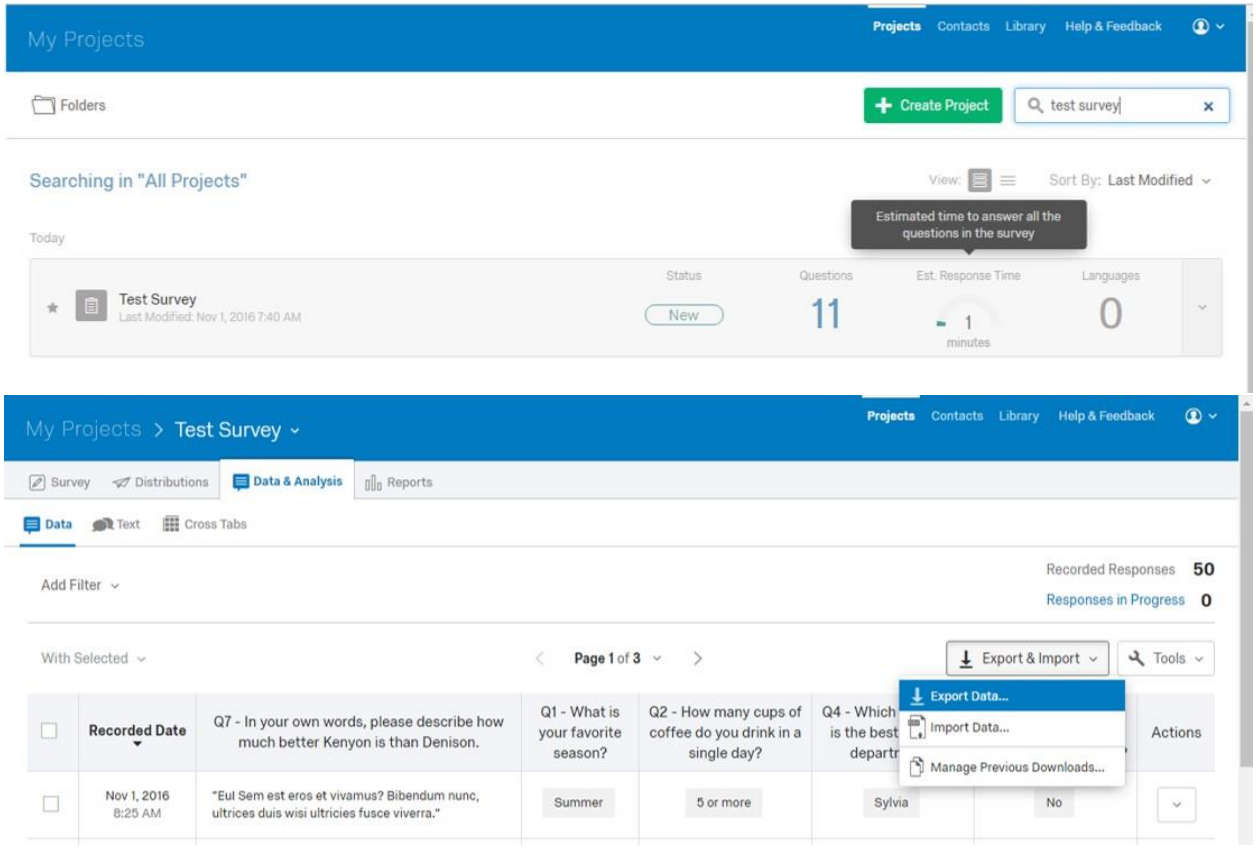
I. Data Set

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.

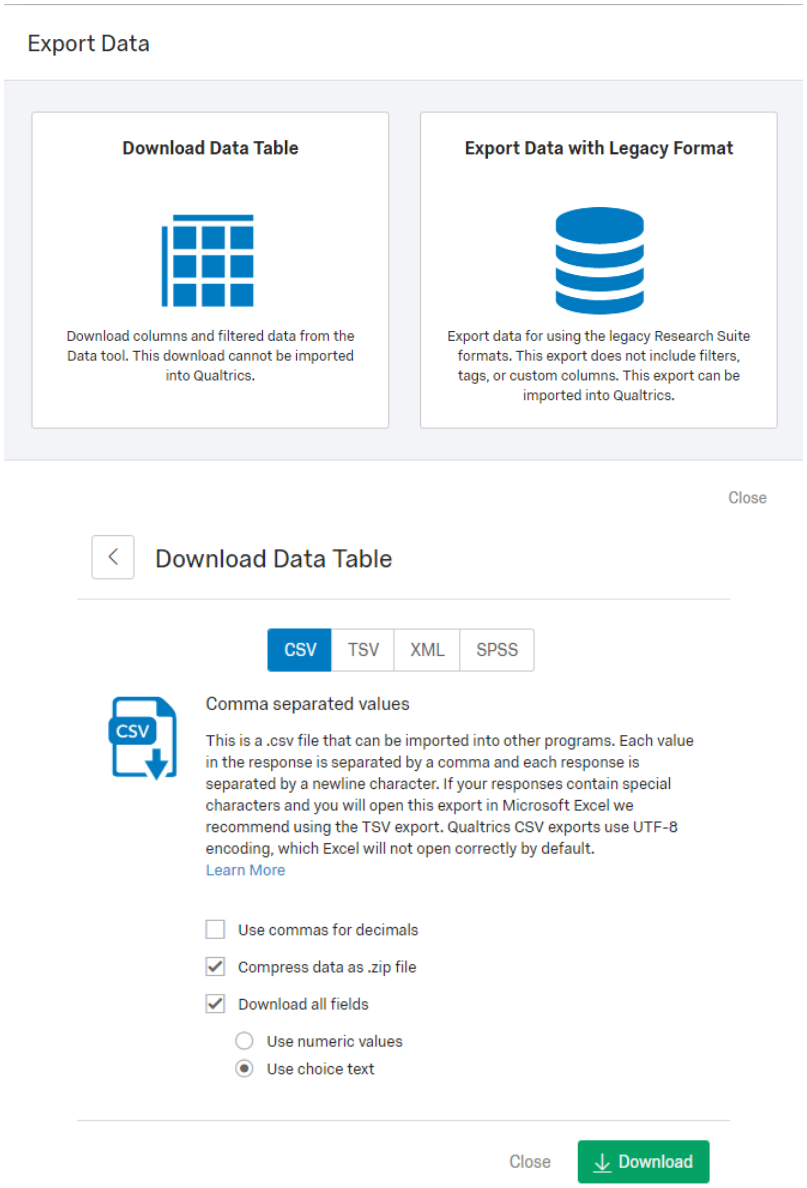
II. Acquiring Your Data Set from Qualtrics

Some 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 (Test 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.

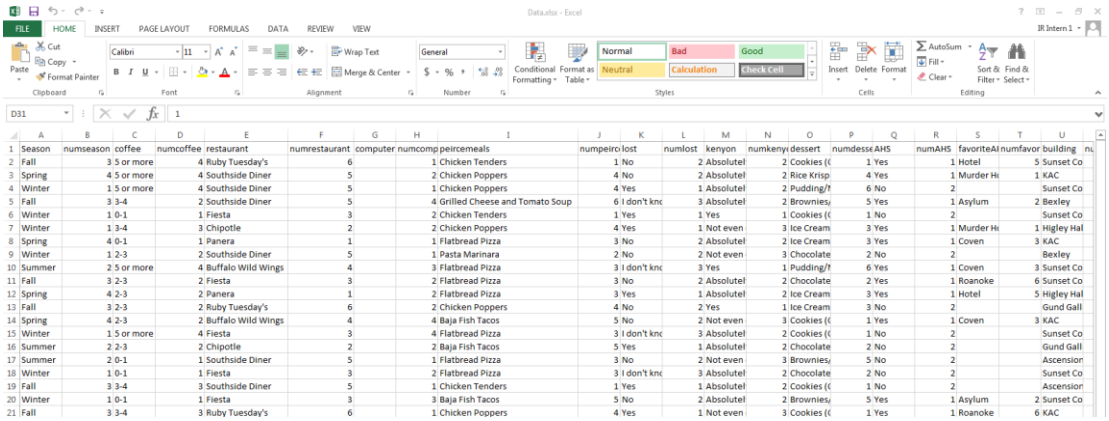


2. After selecting "Export Data", select "Download Data Table". Several options will appear for different downloadable files. This is up to you, but we chose "CSV" to keep the choice text. Checking “Compress data as .zip file” is optional. The output includes when and where the respondent took the survey, how look it took them to take it, progress on the survey, and their answers to survey questions.



3. To get numeric data, go through the same steps as 1 and 2 but instead of selecting “Use choice text” under “Download all fields,” select “Use numeric values.”

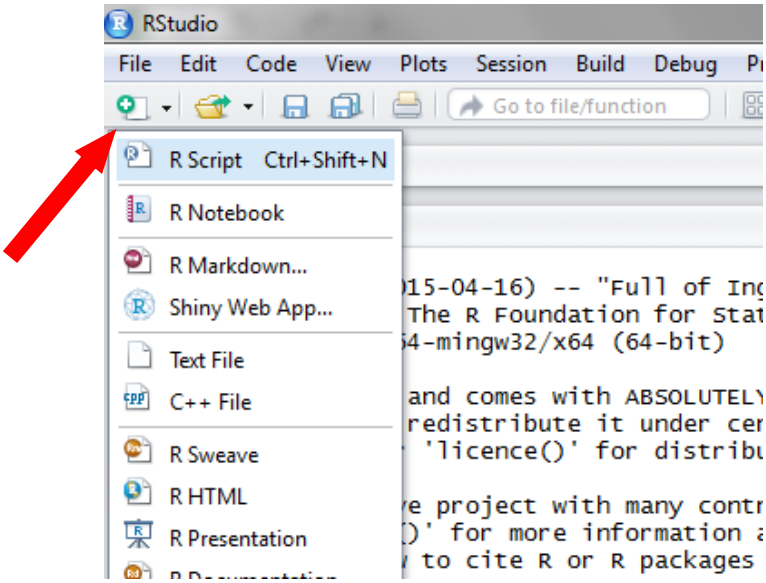
4. We have to manipulate the exported data to use it in R. Therefore, after opening the data in Excel, copy and paste the columns with the survey answers, both numeric and text, into a new Excel file. Rename the columns to the appropriate text variable name (i.e. season, coffee, restaurant, etc.) and numeric values. Save as a .csv file.



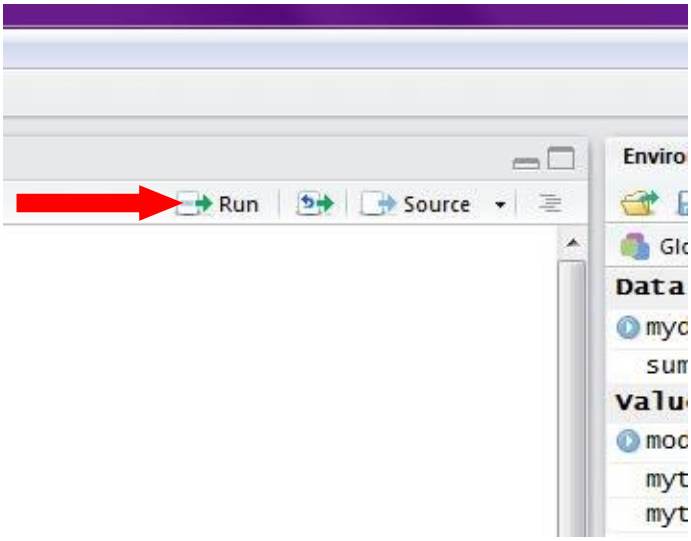
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Season	numseason	coffee	numcoffee	restaurant	numrestaurant	computer	numcomp	peircemeals	numpeirc	lost	numlost	kenyon	numkenyo	dessert	numdesse	AHS	numAHS	favoriteAl	numfavor	building	nu
2	Fall	3	5 or more	4	Ruby Tuesday's	6		1	Chicken Tenders	1	No	2	Absolutel	2	Cookies ((	1	Yes	1	Hotel	5	Sunset Co	
3	Spring	4	5 or more	4	Southside Diner	5		2	Chicken Poppers	4	No	2	Absolutel	2	Rice Krisp	4	Yes	1	Murder Ha	1	KAC	
4	Winter	1	5 or more	4	Southside Diner	5		1	Chicken Poppers	4	Yes	1	Absolutel	2	Pudding/f	6	No	2			Sunset Co	
5	Fall	3	3-4	2	Southside Diner	5		4	Grilled Cheese and Tomato Soup	6	I don't kn	3	Absolutel	2	Brownies/	5	Yes	1	Asylum	2	Bexley	
6	Winter	1	0-1	1	Fiesta	3		2	Chicken Tenders	1	Yes	1	Yes	1	Cookies ((	1	No	2			Sunset Co	
7	Winter	1	3-4	3	Chipotle	2		2	Chicken Poppers	4	Yes	1	Not even	3	Ice Cream	3	Yes	1	Murder Ha	1	Higley Hal	
8	Spring	4	0-1	1	Panera	1		1	Flatbread Pizza	3	No	2	Absolutel	2	Ice Cream	3	Yes	1	Coven	3	KAC	
9	Winter	1	2-3	2	Southside Diner	5		1	Pasta Marinara	2	No	2	Not even	3	Chocolate	2	No	2			Bexley	
10	Summer	2	5 or more	4	Buffalo Wild Wings	4		3	Flatbread Pizza	3	I don't kn	3	Yes	1	Pudding/f	6	Yes	1	Coven	3	Sunset Co	
11	Fall	3	2-3	2	Fiesta	3		2	Flatbread Pizza	3	No	2	Absolutel	2	Chocolate	2	Yes	1	Roanoke	6	Sunset Co	
12	Spring	4	2-3	2	Panera	1		2	Flatbread Pizza	3	Yes	1	Absolutel	2	Ice Cream	3	Yes	1	Hotel	5	Higley Hal	
13	Fall	3	2-3	2	Ruby Tuesday's	6		2	Chicken Poppers	4	No	2	Yes	1	Ice Cream	3	No	2			Gund Gall	
14	Spring	4	2-3	2	Buffalo Wild Wings	4		4	Baja Fish Tacos	5	No	2	Not even	3	Cookies ((	1	Yes	1	Coven	3	KAC	
15	Winter	1	5 or more	4	Fiesta	3		4	Flatbread Pizza	3	I don't kn	3	Absolutel	2	Cookies ((	1	No	2			Sunset Co	
16	Summer	2	2-3	2	Chipotle	2		2	Baja Fish Tacos	5	Yes	1	Absolutel	2	Chocolate	2	No	2			Gund Gall	
17	Summer	2	0-1	1	Southside Diner	5		1	Flatbread Pizza	3	No	2	Not even	3	Brownies/	5	No	2			Ascension	
18	Winter	1	0-1	1	Fiesta	3		2	Flatbread Pizza	3	I don't kn	3	Absolutel	2	Chocolate	2	No	2			Sunset Co	
19	Fall	3	3-4	3	Southside Diner	5		1	Chicken Tenders	1	Yes	1	Absolutel	2	Cookies ((	1	No	2			Ascension	
20	Winter	1	0-1	1	Fiesta	3		3	Baja Fish Tacos	5	No	2	Absolutel	2	Brownies/	5	Yes	1	Asylum	2	Sunset Co	
21	Fall	3	3-4	3	Ruby Tuesday's	6		1	Chicken Poppers	4	Yes	1	Not even	3	Cookies ((	1	Yes	1	Roanoke	6	KAC	

III. R Studio Interface:

When first opening R Studio, you will be presented with a few different windows: Console, Environment/History, and Files/Plots/Packages/Help/Viewer. The console can be used to perform all commands but cannot be saved at the end of a session. However, you can open a new R Script, that can be saved, by either pressing “Ctrl+Shift+N” or navigating to the white box with a green outlined plus in the top left corner and clicking R Script.



When running code in the R Script you can either use ctrl + enter or press the button with the green arrow with the word Run next to it. Doing so runs the line the cursor is currently in or a specific set of lines you have selected. In the Console, you just have to press enter to run the current line.



IV. Inputting Data into R Studio from a .csv

There are few different ways to input data into R from a .csv:

1. One way is to input it interactively. Do this by using the R code: *mydata=read.csv(file=file.choose())* and navigating to your data. Whatever is to the left of the equal sign is what your data is named. You can choose to name it anything. However, if you try to name two things the same name, the second input will override the first.
2. Another way is to use the file extension. Do this by using the R code: *mydata=read.csv("c:\mydata\mydatafile.csv")*

Once your data is read into R, you can choose to "attach" the data. Attaching the data makes the data more accessible with fewer keystrokes. For example, with our data, without attaching *mydata*, in order to run analysis on a specific variable you have to specify what data set the values are coming from by saying *mydata$variablename*. After attaching the data, however, you can just use *variablename*. To attach your data use the code *attach(mydata)*.

V. R Output

At any point when using R Studio, you can ask R for help with a command by using the command *help( ).* The help command provides access to the documentation pages for R functions, data sets, and other objects, both for packages in the standard R distribution and for contributed packages.

1. For a quick overview of your data use the R code "*summary(mydata)*." This outputs the frequency of each response and the minimum, maximum, mean, and quartiles for the numerical values.

2. To obtain the frequency of an individual response, define the variable you want tallied as a table with the code *mytable=table(Season)*. Then to make a table of the frequencies, enter *mytable*. To get the percent of each response, use *prop.table(mytable)*.

3. To compare different variables to one another, define the table of the two variables you want compared as *mytable2=table(Season, coffee).* In a similar way to the individual response, to get the percentage of each response, use *prop.table(mytable2)*.
    a. Another way to do this is to use xtabs. Using xtabs also allows you to have column and row headings. However, doing this requires the library MASS. For more information click here or see 5, below.

4. In order to look at and compare summary statistics for one or more variables, you will have to use the numerical values. You also have to install a package in R. The code below can be run in R to install the package and produce the summary statistics of the three chosen variables. Any variables can be used in their place.

```
install.packages("pastecs")
library(pastecs)
require(pastecs)
sumstat=cbind(numseason, numbuilding, numcoffee)
options(scipen=100) #this forces R to not use scientific notation
options(digits=2) #this makes the output only go to 2 decimal places
stat.desc(sumstat)
stat.desc(sumstat, basic=F) #By setting "basic" to F (false), R removes basic statistics such as number of values, number of null values, number of missing value, minimum and maximum values from the output. For more information about what the output is giving you use the command help(stat.desc)
```

5. To compare two or more variables, use xtabs. To obtain this table use the R code:

```
Install.packages("MASS")
library(MASS)
xtabs(~Season+coffee+AHS)
```

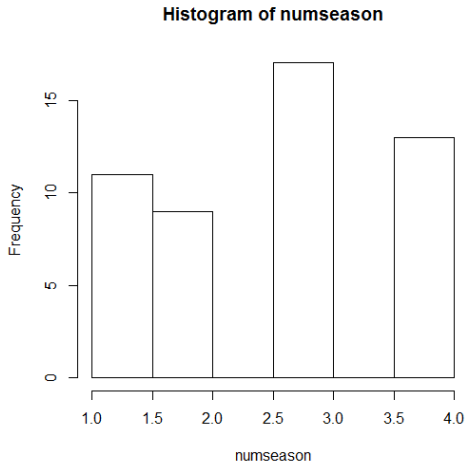
6. To run a regression to examine the relationships between two variables, use the following code:

```
model <- lm(numseason ~ numcoffee) #to create a model of the regression
summary(model) #shows results
```

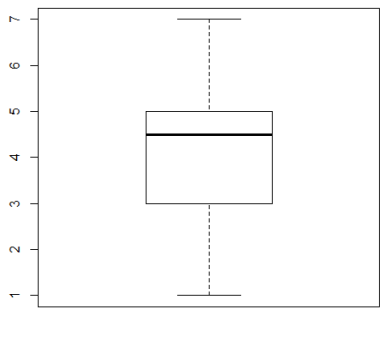
7. For relationships between more than two variables, you can use the same code as in 6 but just add *+variable*. (i.e. model=lm(numseason ~ numcoffee+numAHS) or model=lm(numseason ~ numcoffee+numAHS+numbuilding)) To show results use the same command: *summary(model)*.

Basic Graphics

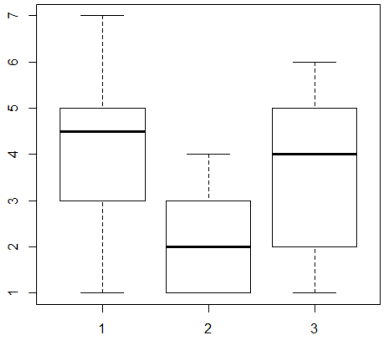
1. To make a histogram, use the code *hist( )*. For example, *hist(numseason)* produces this graph:



2. To make a boxplot, use the code *boxplot()*. You can view multiple boxplots on one graph by adding more variables separated by a comma. For example, *boxplot(numbuilding)* produces boxplot 1, whereas *boxplot(numbuilding,numcoffee,numrestaurant)* produces boxplot 2.



Boxplot 1



Boxplot 2

3. To make a scatterplot a single variable on an x,y axis, use the code *plot()*. For example, you could plot *numbuilding* using *plot(numbuilding)*.
4. To make a scatterplot two variable (x and y), use the code *plot(x,y).* Where *x* is the variable on the x-axis and *y* is the variable on the y-axis.