# Prep Exercise (PE09) Associated Rules Mining

### General Instructions

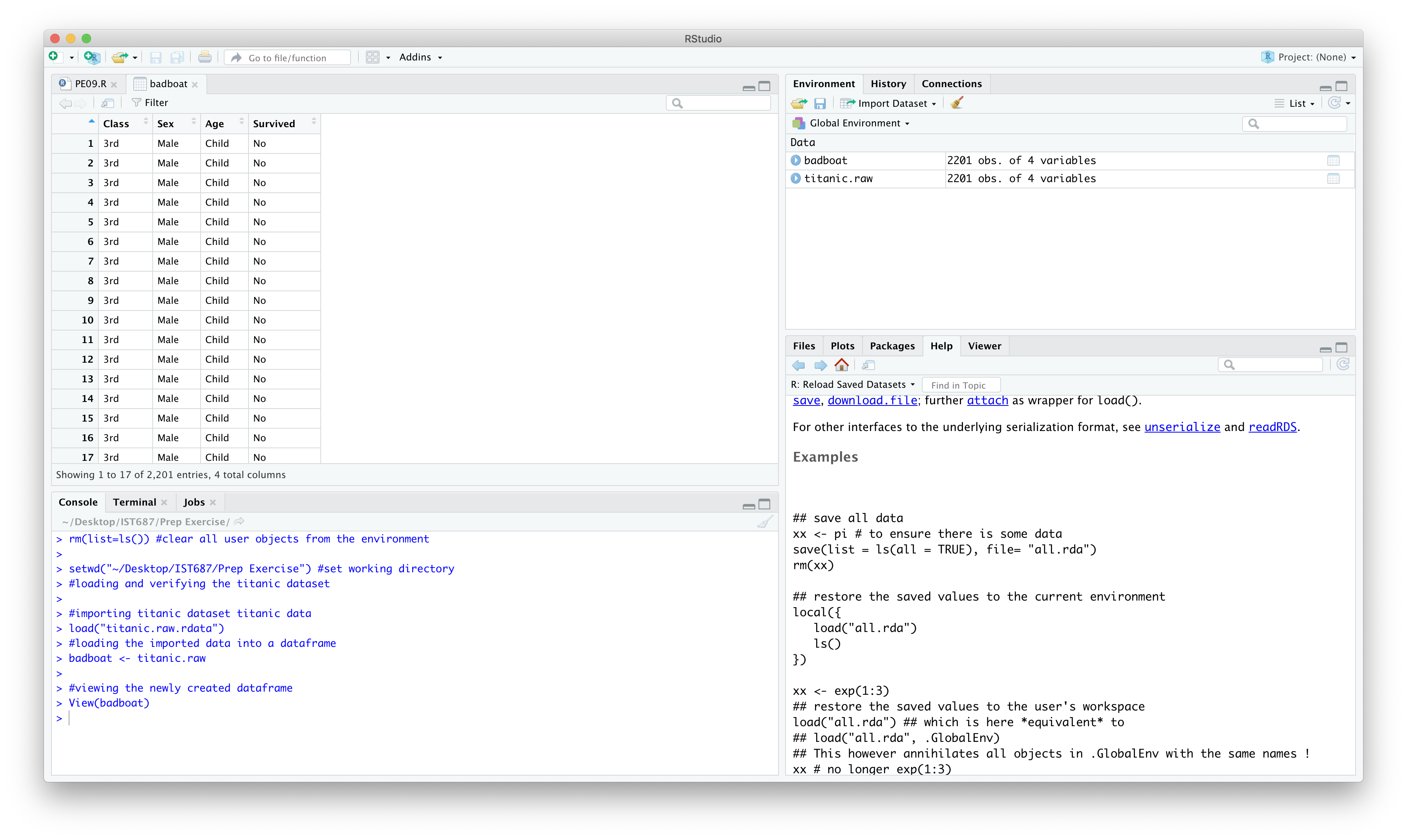
1. For this exercise you will answer all of the questions in this document and turn it in to Blackboard.
2. Before you get started make sure to read Chapter 17 of *An Introduction to Data Science* and execute the code throughout the chapter to gain familiarity.
3. Getting Started:
   1. Up until now we have done relatively simple data exploration and modelling, here is where we enter the territory of machine learning. This week we will be examining associated rules mining, a form of machine learning that discovers relationships between variables within large or small datasets. A practical example of associated rules mining is seen in the way grocery stores operate, maximizing their profits by organizing their stores in a manner that maximizes the probability that a person who purchases “x” will also purchase “y”. However, in this PE and homework we will be examining associations within a dataset containing titanic data.
   2. As usual we will use this Prep Ex to set you up for the homework exercises and test your knowledge of materials within the chapter reading. Let’s begin…

### Prep Exercise

1. **Getting Ready: Loading and Verifying the Titanic Dataset**
   1. An R dataset containing the titanic data is available on the Blackboard site. Download it to your computer and use the *load()* command to bring it into your RStudio environment.
   2. If you are having trouble using the *load()* command, make sure you downloaded the titanic data to your working directory. You can check what is your working directory in R using the code below:

**getwd()**

* 1. Load the imported data into a dataframe called “badboat”.
  2. Run the *View()* command to verify that the dataframe has been loaded correctly and place a screenshot of the dataset below (do not worry about screenshotting the entire dataframe, a portion will do).

****

1. **Understanding Terminology that will be used in this PE and HW.**
   1. In a paragraph or two explain the concept of a sparse matrix and how to identify whether or not data is a sparse matrix.

**Sparse matrix is a 2-dimentional array with 0 and 1 as the array elements. It contains majority of zeros and fewer ones. It is called sparse as the matrix contains very few ones. If a data element is present, then its presence is indicated by 1. If a data element is not present, then it is indicated by 0.**

**Sparsity is number of zeros by total number of elements in the matrix. A matrix is sparse if its sparsity is greater than 0.5**

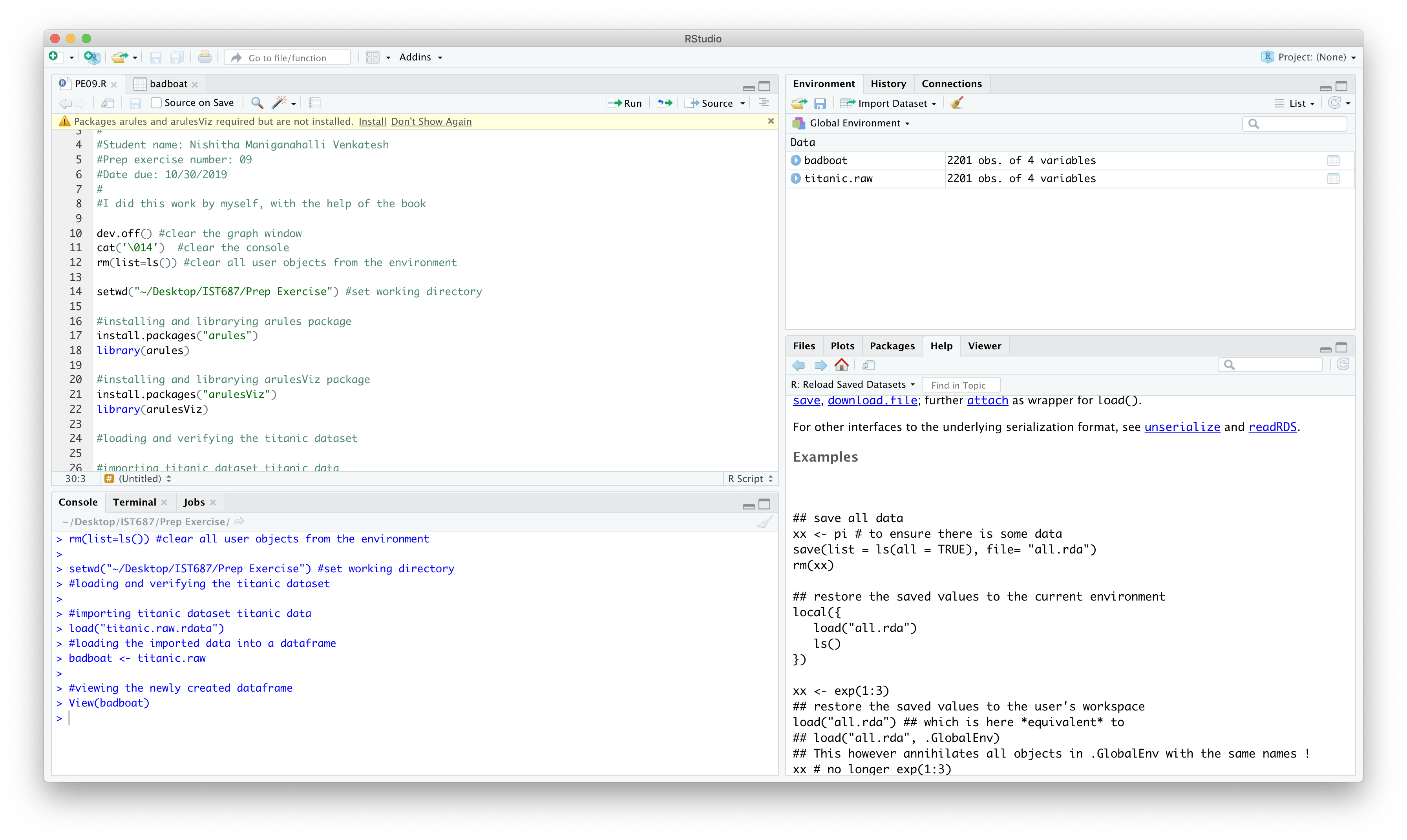
* 1. In a few sentences explain what a dense matrix is and its defining feature.

**Dense matrix is the one with most number of non-zero elements. Density of a matrix is 1 minus sparsity.**

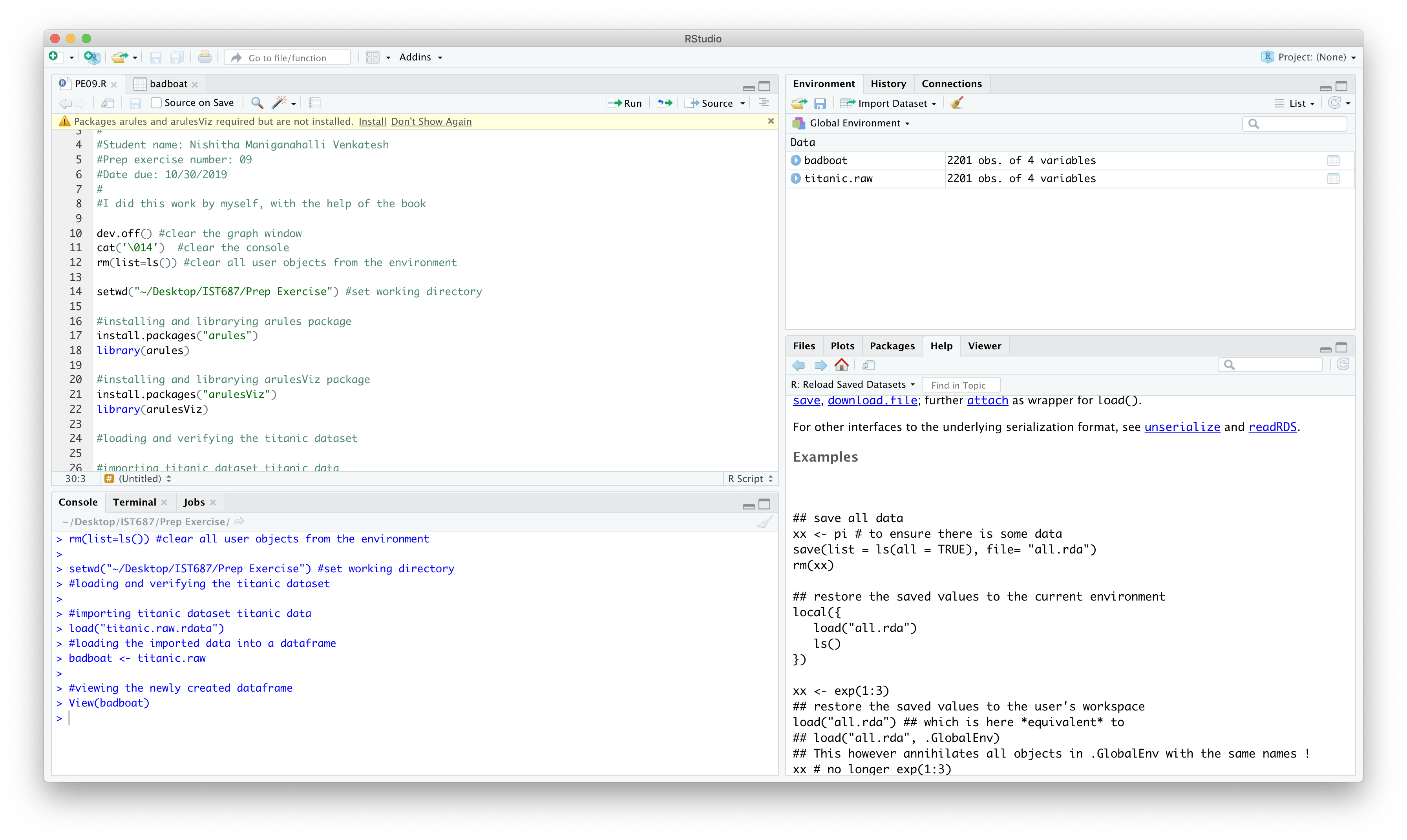
* 1. Explain contingency tables and how you would go about creating one in RStudio (if needed, you can google “contingency table in r”, there are many resources on the web). How would you go about reading /understanding a contingency table?

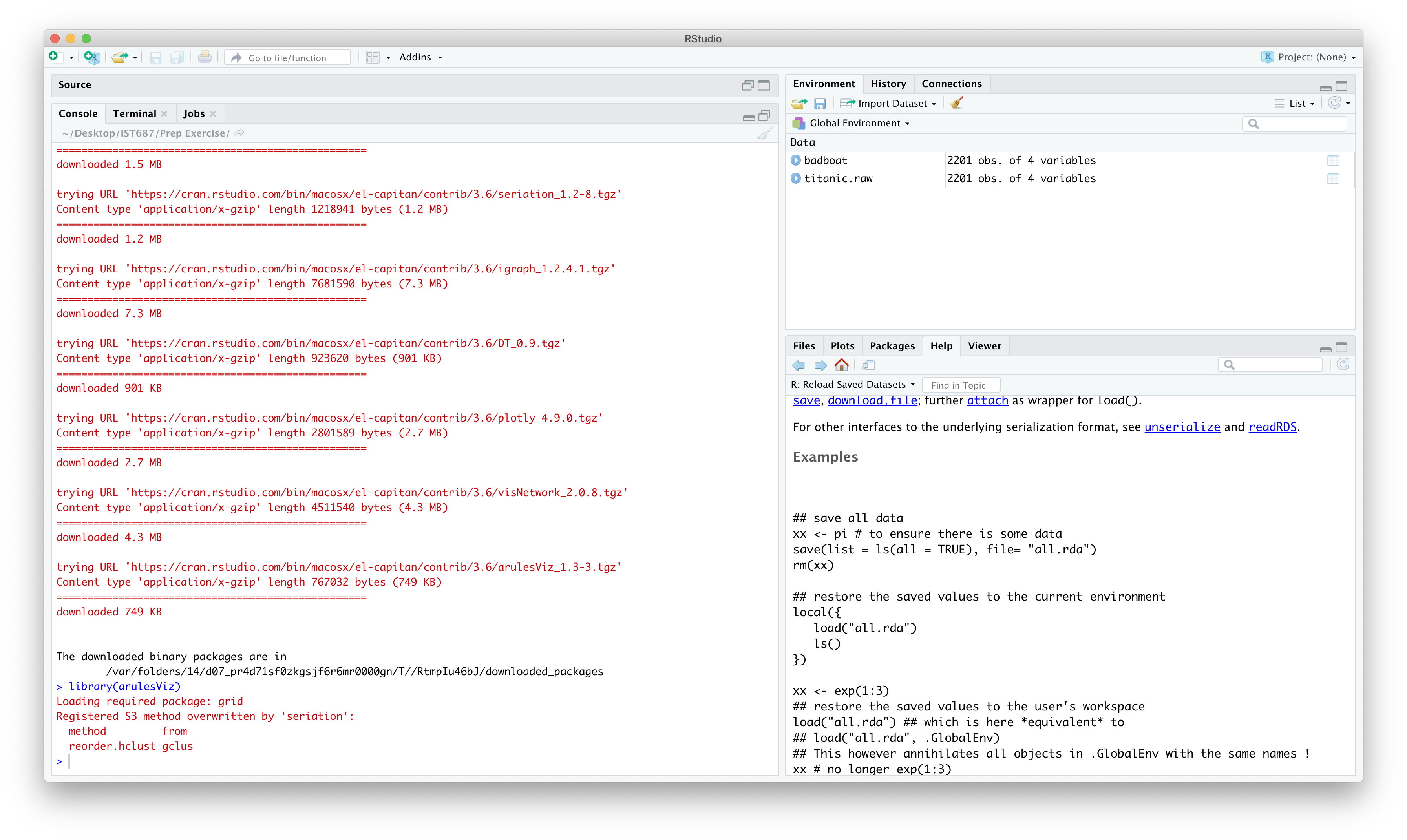
**Contingency table is a frequency distribution table which shows the relation between two variables. In RStudio, contingency table can be created for vectors using the table() command, where the first argument is for row and second is for column. The transpose of the table can be obtained using t() command. The contingency table determines how the row and column variables interact.**

1. **Loading Necessary Packages.**
   1. In the homework portion of this week’s lesson you will need two packages; *arules* and *arulesViz*.
   2. Install both packages and check/update them using the *library()* command to ensure the packages exist. Paste a screenshot of the code and respective output below.

****

****

****

****

1. **List any additional resources that you used here.**

<https://www.datasciencecentral.com/profiles/blogs/contingency-tables-in-r>

1. **Be sure to save your R file as this will become the starting code for your homework.**

***You must submit all Prep Exercises to blackboard prior to the deadline specified for each assignment.*** PE assignments are due on the evening prior to the lecture class. Late PE assignments will not be accepted for credit.

*Materials provided by the instructors of this course are copyrighted 2018 by Jeffrey Saltz and Jeffrey Stanton, or by the respective instructor that produced them. These materials are provided to enrolled students for their exclusive use during the semester of instruction. Uploading instructor-produced materials to Internet sites (e.g., Coursehero) without permission is considered a breach of professional ethics and will be reported as a possible academic integrity violation.*